Language acquisition with communication between learners

Rasmus Ibsen-Jensen[1,†,*], Josef Tkadlec[1,†], Krishnendu Chatterjee[1], and Martin A. Nowak[2]

[1]IST Austria, Klosterneuburg, A-3400, Austria

[2]Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, MA 02138, USA

[†]These authors contributed equally to this work.

[*]Corresponding author: ribsen@ist.ac.at

We consider a class of students learning a language from a teacher. The situation can be interpreted as a group of child learners receiving input from the linguistic environment. The teacher provides sample sentences. The students try to learn the grammar of the teacher. In addition to just listening to the teacher, the students can also communicate with each other. The students hold hypotheses about the grammar and change them if they receive counter evidence. The process stops when all students have converged to the correct grammar. We study how the time to convergence depends on the structure of the class room by introducing and evaluating various complexity measures. We find that structured communication between students, although potentially introducing confusion, can greatly reduce some of the complexity measures. Our theory can also be interpreted as applying to the scientific process, where nature is the teacher and the scientists are the students.

Keywords: Language learning; Inductive Inference; Population structures in learning

1 Introduction

In traditional language learning theory [1, 2, 3], there is a teacher and a learner [4, 5, 6]. The teacher uses a particular grammar and provides sample sentences from the corresponding language. A language is a set of finitely or infinitely many sentences. A grammar is a finite list of rules that specifies the language. The learner has a search space of candidate grammars. The task for the learner is to converge to the grammar of the teacher after having heard a sufficient number of sentences. This setting for learning is called "inductive inference" [7, 8]. The goal is to infer the underlying rules from examples. The teacher cannot directly communicate the rules of the grammar, (s)he only provides sample sentences consistent with it.

Learning by inductive inference is more general than natural language acquisition. It arises whenever generative rules are supposed to be inferred from examples. It is the basis for mutual understanding in human communication. It is also the activity of scientists searching for the laws of nature [9]. The scientists conduct experiments and the nature gives the answers. Then the scientists seek to formulate the underlying rules, the grammar of nature. In the present work, we focus on language learning as a particular case of cultural transmission.

Learning theory is often concerned with positive or negative results about the learnability of sets of grammars [10, 11, 12, 13, 14, 15, 16, 17]. It is the basis for a mathematical formalization of what Chomsky calls "universal grammar" [8, 18, 19]. Several works also considered the computational problems related to learning [20, 21, 22]. In the evolutionary dynamics of human language acquisition, the question is extended to asking under which conditions a population of speakers learning from each other can converge to a coherent language [23, 24, 25, 26, 27].

In this paper we explore a new setting. There is a teacher (either a person, or a body of knowledge, or the linguistic environment or nature) and a population of learners. In addition to just listening to the teacher, the learners can also communicate with each other. At each moment, each learner holds a hypothesis as for what is the teacher's grammar and can update this hypothesis upon hearing a single sentence from the teacher or some other learner. The learners and the teacher speak and listen to one another until, eventually, all learners successfully learn the grammar of the teacher. In the next section we introduce a model in which the communication among learners and the teacher proceeds in an organised way. We study which communication structures improve – or obstruct – the efficiency of this learning process.

The efficiency of the learning process also depends on the power of individual learners. Here we consider learners of two different types: weak memoryless learners and powerful batch learners. As far as memory is concerned, these two types of learners serve as a lower and upper bound for human learning capacity [5, Section 13.3.3]. Memoryless learners hold, at any moment, a candidate grammar. Whenever they receive a counterexample (a sentence that doesn't belong to the language corresponding with their current grammar) they randomly choose another grammar from their search space. They are called "memoryless" because they could pick a grammar which they have already rejected. In contrast, batch learners keep track of all the inputs they have received so far and for their hypothesis they always select grammar that is most consistent with the sentences they have observed so far. When learning from a single teacher without other inputs, both types of learners have the property of consistency: once they find the right grammar they do not change it anymore.

The underlying dynamical system can be seen as a new kind of evolutionary process. Candidate grammars spread in the population of learners. The teacher, or the environment, selects for particular grammars. The process stops when all learners have adopted the correct grammar. The basic question is: How is the time to linguistic coherence affected by the population structure?

2 Model

In this section, we first introduce a general model for language learning with structured communication between learners. Next we present two types of learners (memoryless (p, q)-learners and powerful batch learners) that we later analyze in detail. Finally, we introduce a complexity measure called *rounds complexity* that we use to evaluate the efficiency of the learning process for different communication structures and types of learners. Our main scientific finding is as follows: while communication between learners can potentially cause confusion and certain communication structures between learners indeed do slow down the learning process, we present communication structures that can significantly expedite the learning process.

The process of learning a language can be modelled in a variety of ways [28, 29, 30, 31, 32, 33]. In the traditional setting there is a single teacher and a single learner, and only the teacher communicates with the learner. Here we extend the traditional setting as follows:

1. We consider a single teacher and a population of learners.

2. The population of learners can communicate among each other.

3. We consider structured communication between the learners and study whether such communication can improve the efficiency of the process.

For clarity of presentation, we identify a grammar (a list of rules) with the language (a set of sentences) it generates. The hypothesis of each individual at each time is thus a language. (Recall that the units passed at each communication event are sentences.)

Single learner. In the traditional "single teacher – single learner" scenario, the teacher speaks some language L_1 unknown to the learner and repeatedly generates sentences from L_1. The learner has a search space of possible languages $L_1, L_2, \ldots$ and initially holds an arbitrary hypothesis as for what the teacher's language is. Upon hearing each sentence from the teacher, the learner can update this hypothesis. The process ends when the learner's hypothesis becomes L_1.

Structured learning for multiple learners. In our case, there is a group of $n + 1$ individuals (one teacher and n learners). There is a set L of ℓ languages $L_1, \ldots, L_\ell$. Each language consists of sentences (one sentence can belong to multiple languages).

The communication structure among learners is represented by a directed graph (network) where nodes correspond to individuals (including the teacher) and an edge (arrow) from individual A to B means that A listens to B. At each moment, each learner holds a hypothesis $L_i \in L$ regarding what the teacher's language is. Initially, teacher holds L_1 and the hypotheses of the learners are arbitrary. In every round of the learning process we pick all the edges of the graph one by one, in random order. Every time an edge is picked, the speaker of that edge generates a sentence from the language she is currently hypothesizing and the listener of the edge can update his hypothesis. The process stops when all the learners learned the teacher's language L_1.

Example. As a toy example, consider a single teacher T and two learners A (Alice) and B (Bob) such that both A and B listen to T and moreover B listens to A. Suppose that there are two languages L_1, L_2 that don't overlap at all. Suppose that A's initial hypothesis is L_2 while B starts with L_1 (T starts with L_1 too). Finally, suppose that both learners follow the same simple update rule: whenever they hear a sentence they can not parse, they switch their hypothesis to the other possible language with probability 80 % (and keep it otherwise).

In this example, a single round can play out as follows (see Figure 1(b)): First we pick the edge between B and T. B receives a sentence he understands and keeps his hypothesis L_1. Next we

pick the edge between B and A. B receives a sentence from A's language L_2. He can't parse it and (with probability 80 %) he switches his hypothesis to L_2. Finally, we pick the edge between A and T. A receives a sentence she can't parse, still (with probability 20 %) she sticks to her current hypothesis L_2. As an outcome of the round, both A and B now hold the wrong hypothesis L_2.

Note that had we first picked the edge between A and T, A could have switched to L_1 with probability 80 % and the whole process would have finished in a single round. Allowing learners to speak among themselves can create confusion and can result in less efficient learning.

Memoryless learners: (p, q)-learning. Here we describe a type of a memoryless learner that we call a (p, q)-learner. There are two positive numbers $p, q \in [0, 1]$ with $p + q \leq 1$. Upon hearing a sentence, a (p, q)-learner updates her hypothesis as follows: (a) if the learner holds the same language as the speaker, then nothing changes; (b) if the learner holds a different language from the speaker, then:

1. with probability p the learner's hypothesis changes to the language of the speaker;
2. with probability q the learner's hypothesis does not change;
3. with probability $(1-p-q)/(\ell-2)$ the learner switches to one of the remaining languages (i.e., with the remaining probability one of the other languages is chosen uniformly at random).

An illustration is presented in Figure 1(a).

The parameters p, q can model various features of language learning. (a) The parameter q can represent the overlap between different languages, such that even if the languages of the speaker and the listener are different, the sentence from the speaker can be parsed by the listener and hence the listener does not switch. (b) The parameter p represents the bias to switch to the language of the speaker by listening to a single sentence. Note that since the switch happens by listening to a single sentence we consider that p is proportional to $1/\ell$.

Discussion of (p, q)-learners. We explain how our model of a (p, q)-learner generalises several classical language learning scenarios considered in the literature.

- *RWA:* A model of random walk (without greediness and single-value constraints) (RWA) on languages has been considered in [6, Section 4.2.1] where if the speaker and the listener have different languages, then the switch is uniformly at random among all languages. In the above setting we achieve this with $p = q = 1/\ell$.

- *SS:* A model of language learning with symmetric language overlap (SS) was considered in [5, Section 13.3.2]. The overlap was characterised by parameter a in [5, Eqn. (13.26)], which precisely corresponds to the parameter q in our model.

- A speaker can speak sentences that are either helpful or hindering to learning. For example, with helpful sentences, the switching probability p can increase to c/ℓ, where $c > 1$. In contrast, with hindering sentences, it can decrease to c/ℓ, where $c < 1$.

- Another aspect in communication that has been considered in [6, Section 3.3] is the presence of noise. Due to the presence of noise, the sentence from a speaker might not be received by a listener, and hence the listener does not switch. The parameter q in our model can represent such noise in the communication.

The symmetry (SS) generalises (RWA) with overlap between languages. RWA and SS represent the simplest examples of language learning. Extension to the case of non-symmetrically overlapping languages is discussed in Supplementary Information (SI) Section 3.7.

Batch learners. The other type of the learner we consider is a powerful batch learner. A batch learner remembers all the inputs she received so far and for her hypothesis, she always selects the language that is most consistent with all her observations (initially, her memory is empty). More formally, having observed sentences $s_1, s_2, \ldots, s_n$, the batch learner updates her hypothesis to a language L_i from her search space for which the size of the set $L_i \cap \{s_1, s_2, \ldots, s_n\}$ is maximised. We consider batch learning in the case of symmetric language overlap $q < 1$. That is, the size of the overlap of any k languages is equal to q^{k-1} times the size of any of the languages (see SI Section 2.2 for details).

The main scientific question: Rounds complexity. While a basic question in learning theory is about identification of the correct language in the limit, an equally important question is about the efficiency of the learning process, which has been described in details in [21, Chapter 2]. The efficiency of the learning process is determined by the speed of convergence to the correct language by the whole population. The main scientific question we investigate in this work is the effect of communication structures in the learning process. More precisely, we are interested in communication structures that speed up the learning process. In order to asses the efficiency of the process, we compute the expected (average) number of rounds until the process has converged (that is, all

learners learned the teacher's language). We refer to this as the *rounds complexity* of the process. We discuss other relevant measures later.

Illustration of the scientific question. We illustrate our scientific question on a small example with four learners for RWA learning model of [6, Section 4.2.1]. As baseline we consider that there is no communication between the learners (denoted as the empty graph). We illustrate four possible communication structures in Figure 2. We observe that with respect to the expected number of rounds the communication structures Graph B and Graph C are worse than the empty graph, whereas the communication structure Graph D is better than the empty graph. The main take away message is: while some communication structures are worse for the learning process, others can lead to more efficient learning.

3 Results

Remember that n is the number of learners. We present both theoretical results and simulation results. In theoretical results we introduce several communication structures (empty graph, complete graph, tree graph, Layered Hierarchy graphs). For each communication structure we analyze the rounds complexity (i.e. the expected number of rounds until all individuals have learned teacher's language). Then we compare the rounds complexities in the limit of large n. Later we show matching numerical simulations for small n.

Our theoretical results are presented in terms of n and T, where T denotes the expected number of rounds in the single teacher and single learner case (T also corresponds to the sample complexity of [22]). For example, in case of single learner and RWA or SS with ℓ languages we have $T \approx c \cdot \ell$ for some constant $c > 0$. First we consider (p, q)-learners.

Remark on asymptotic complexity. When comparing the rounds complexity of two processes A and B in the limit of large population size n, the improvement can be either a *constant-factor* if the dependency on n is the same (e.g. $A = 10 \cdot n$ vs. $B = 5 \cdot n$), or *asymptotic* if the dependency on n is different (e.g. $A = 10 \cdot n$ vs. $B = 10 \cdot \sqrt{n}$). In the former case we say that the asymptotic complexities match. In the latter case we say that B has better asymptotic complexity than A (expression $\sqrt{n}$ is much smaller than n for large n). For detailed treatment see [34, Section 1.3]

Classroom teaching: empty graph (Figure 3(a)). For the baseline comparison we consider the most

natural extension of the single learner scenario: The empty graph consists of multiple learners who all listen to the same teacher and don't communicate among each other at all.

The rounds complexity is at most $c_1 \cdot T \cdot \log n$, where $c_1 > 0$ is a constant (see SI Section 3.2). Hence the rounds complexity is linear in T and logarithmic in n. In particular, for RWA and SS, the upper bound is $c_1 \cdot \ell \cdot \log n$. Moreover, for RWA and SS, we provide matching lower bounds to show that the upper bound is optimal, and hence the upper bound cannot be improved in general.

Complete graph (Figure 3(b)). The opposite extreme is the complete graph where all learners speak to each other. Even in the simplest RWA and SS models, the complete graph has rounds complexity that is exponential in n (see SI Section 3.4). Hence it is extremely inefficient for the learning process and we will not discuss complete graphs further.

Tree graph (Figure 3(c)). Speaking to many other individuals is more demanding for the speaker. If we insist that every individual speaks to only a constant number of other individuals, we naturally obtain a tree graph. In terms of rounds complexity, the tree graph is worse than the empty graph but only by a constant factor (not asymptotically).

For simplicity we consider the binary tree (every individual speaks to at most two others). The vertices are organised in levels, and the teacher has level 0. Every vertex at level i has at most two incoming edges from vertices of level $i + 1$, and each vertex (other than the teacher) has exactly one outgoing edge. Vertices without incoming edges are called leaves. For every n, we construct a binary tree which has at most $\log n$ levels. We show that the rounds complexity is at most $c_2 \cdot T \cdot \log n$, where $c_2 > 0$ is a constant (see SI Section 3.5). Hence, as for the empty graph, the dependency is linear in T and logarithmic in n. The constant c_2 is greater than c_1, and thus the tree is worse than the empty graph by a constant factor, although asymptotic complexities are the same. Moreover, for RWA and SS, we establish similar lower bounds as in the case of empty graph.

Layered Hierarchies. Our most interesting results are related to certain hierarchical structures that we call *Layered Hierarchies*. We show that certain Layered Hierarchies might improve the rounds complexity, but do not improve the asymptotic complexity, whereas Layered Hierarchies with quickly growing group sizes improve even the asymptotic complexity.

Description of Layered Hierarchies (Figure 3(d),(e)). We start with a general description of Layered Hierarchies. In a k-Layered Hierarchy graph the learners are partitioned into groups (or layers) $S_1, S_2, \ldots, S_k$. The edges go from each group S_i to the previous group S_{i-1}, for $2 \leq i \leq k$, and

the edges from the first group S_1 go to the teacher. An illustration of 2-Layered Hierarchy and k-Layered Hierarchy graphs are shown in Figure 3(d),(e), respectively. Incidentally, the empty graph can be called the 1-Hierarchy. We have described the principle of Layered Hierarchy graphs without specifying the sizes of the groups which we discuss below.

"Slowly growing" Layered Hierarchies. The group sizes can be of various types, and we discuss the simple ones below: (a) *Constant size.* All group sizes are the same. (b) *Additive growth.* The next group size is a constant more than the current group size. (c) *Multiplicative growth.* The next group size is a constant times larger than the current group size. Let us consider the above group sizes for three layers ($k = 3$).

- *Constant size.* In this case, each group has $n/3$ learners. In particular the first group has $n/3$ learners, and even just considering the time to convergence for the first group, in general the rounds complexity is at least $c_1 \cdot T \cdot \log(n/3)$. Thus the asymptotic complexity does not change with respect to the empty graph.

- *Additive growth.* Let the group sizes be x, $2 \cdot x$, and $3 \cdot x$. Since the sum of the group sizes is n, the first group size is $n/6$. Similarly, to the above item, in general the rounds complexity is at least $c_1 \cdot T \cdot \log(n/6)$. Again the asymptotic complexity does not change with respect to the empty graph.

- *Multiplicative growth.* Let the group sizes be x, x^2, x^3. Since the sum of the group sizes is n, the first group size is $x \approx n^{1/3}$, and similarly to the previous items, in general the rounds complexity is at least $c_1 \cdot T \cdot \log n^{1/3} = \frac{1}{3} \cdot c_1 \cdot T \cdot \log n$. We observe even in this case the asymptotic complexity does not change as compared to the empty graph.

We remark that even though the asymptotic complexity doesn't change, the rounds complexity of Layered Hierarchies is in practice often smaller than that of an empty graph by a constant factor. The corresponding simulation results are presented in SI Section 5.2 (Figure SI.3).

Exponentially growing Layered Hierarchy. We now consider Layered Hierarchy graphs where the group sizes grow exponentially, and show that they provide a significant asymptotic improvement over the empty graph among learners. We start with the simpler case of Exponential 2-Layered Hierarchy (for brevity 2-Hierarchy in the sequel), then describe the general case of Exponential Layered Hierarchy (for brevity, Hierarchy). In the 2-Hierarchy, intuitively, the teacher quickly

teaches a small group of learners and then uses them as additional teachers to speed up the teaching of the rest of the population. The Hierarchy iterates this construction. The precise descriptions are as follows:

- *2-Hierarchy.* We split the learners into two groups S_1, S_2, where the size of S_1 is proportional to $\log n$, which is written as $|S_1| \propto \log n$. The graph then consists of all the edges from S_1 to the teacher and all the edges from S_2 to S_1; see Figure 3(d) with $|S_1| \propto \log n$ and $|S_2| \propto n$. For example, a 2-Hierarchy of 1 000 learners has $|S_1| = 10$ and $|S_2| = 990$.

- *Hierarchy.* Hierarchy is obtained by iterating the construction of the 2-Hierarchy. We split the learners into groups $S_1, \ldots, S_k$ such that the first group consists of 2 learners and that each following group is exponentially larger than the previous group: $|S_{i+1}| \propto 2^{|S_i|}$. The edges go from each group to the previous group and from the first group to the teacher; see Figure 3(e) with $|S_1| = 2$ and $|S_{i+1}| \propto 2^{|S_i|}$ for $i = 1, \ldots, k-1$. A Hierarchy of 1 000 learners would include 2, 4, 16, and 978 learners in the respective groups.

We establish the following results (see SI Section 3.6).

- For the 2-Hierarchy the expected number of rounds is at most $c_3 \cdot T \cdot \log\log n$, where $c_3 > 0$ is a constant. While the rounds complexity dependency is linear in T, the dependency is double logarithmic in n, which is significantly better than logarithmic. Moreover, even if we interpret dependency in T, for large n, we have $c_1 \cdot \log n > c_3 \cdot \log\log n$. Thus, for a reasonably large population the 2-Hierarchy is better than the empty graph.

- For Hierarchy we show the expected number of rounds is at most $c_4 \cdot T \cdot \log^\star n$, where $c_4 > 0$ is a constant and $\log^\star$ ("log star") is the iterated logarithm, which is a *very* slowly increasing function that appears in many computer science applications. Formally, $\log^\star n$ is the number of times the logarithm function must be iteratively applied to number n before the result is less than or equal to 1. For any $1 \leq n \leq 2^{256} \sim 10^{77}$ we have $1 \leq \log^\star n \leq 4$, and thus $\log^\star(n)$ is effectively constant for all practical purposes. The Hierarchy therefore provides dramatic improvements over the empty graph.

For 2-Hierarchy we again provide matching lower bounds for RWA and SS to show that the upper bound cannot be improved in general.

Remark on rounds complexity. If we compare the empty graph and the 2-Hierarchy for RWA or SS, where the number of languages is finite and equal to ℓ, for memoryless learners we obtain that the rounds complexity is proportional to $\log n \cdot \ell$ for empty graph, and proportional to $\log \log n \cdot \ell$ for 2-Hierarchy. Note that our results establish how the population structure influences the dependency on n. The improvement of $\log n$ to $\log \log n$ can be significant when ℓ is large. For example, if $n = 16$, then $\log n$ is 4 whereas $\log \log n$ is 2. Hence the rounds complexity decreases from 4ℓ to 2ℓ, which can be significant speedup in practice.

Other complexity measures. The expected number of rounds (i.e. rounds complexity) is the most natural measure for the efficiency of the learning process. However, there are other relevant measures which we discuss now.

1. The *communication complexity* is the expected number of communication events until the process converges. Each communication event represents one usage of one edge in the graph. The measure represents the total amount of sentences that need to be exchanged in the whole population.

2. The *bottleneck complexity* is the expected maximum number of communication events that need to be done by a single individual, which could be the teacher or one of the learners, until the process converges. If the bottleneck is the teacher then this measure relates to the amount of sentences that need to be extracted from the environment.

Relevance of the complexity measures. In distributed computing and network computation, rounds complexity is a very relevant notion, and communication complexity (or message complexity) is also well-studied [35, 36]. Typically, in distributed computing the communication structures are symmetric and bottleneck is not widely studied, however in hierarchical network structures, bottleneck is an important complexity measure [37]. This work shows that these complexity measures from network theory become relevant for language learning in population structures, and in particular, the population structure can affect the complexity measures.

Results for other complexity measures. We now present our results for the other complexity measures for the graphs we consider. We first note the following:

1. *Communication complexity.* The communication complexity is always the rounds complexity times the number of edges in the graph (including the edges to the teacher).

2. *Bottleneck complexity.* The bottleneck complexity is always the rounds complexity times the max-degree of the graph.

We show that the empty graph is optimal with respect to communication complexity (see SI Section 3.3). There is no graph that can be better than the empty graph for the communication complexity. The bounds for communication and bottleneck complexity for all the graphs are obtained from our results on rounds complexity. Note that the asymptotic communication complexity has the same dependency on T and n in all cases except for the complete graph. However, the associated constants are different, with the empty graph having the least constant among them. All the results are presented in Table 1.

Discussion of the results for (p, q)-learners. As mentioned above, the empty graph is optimal with respect to the communication complexity. The complete graph is worse in terms of all complexity measures. The tree graph matches the asymptotic complexity of the empty graph with respect to communication and rounds complexity, and improves the bottleneck complexity from $n \log n$ to $\log n$. The 2-Hierarchy matches the asymptotic complexity of the empty graph with respect to communication complexity, significantly improves the round complexity dependency from $\log n$ to $\log \log n$ and improves the bottleneck complexity from $n \log n$ to $n \log \log n$. The Hierarchy matches the asymptotic communication complexity of the empty graph and significantly improves the round complexity from $\log n$ to $\log^\star n$ and the bottleneck complexity from $n \log n$ to $n \log^\star n$.

Results for batch learners. For batch learners under the assumption of symmetrically overlapping languages we obtain results that are similar in spirit to those for (p, q)-learners. The complete graph is much worse than the empty graph in terms of all complexity measures. The tree graph improves the bottleneck complexity as compared to the empty graph. The 2-Hierarchy graph improves both the rounds complexity and the bottleneck complexity as compared to the empty graph. The results are summarised in Table 1 (see SI Section 4 for details).

Numerical simulations (Figure 4). Our theoretical results establish asymptotic complexity bounds that apply in the limit of large population sizes. To complement them, we present numerical simulations for small population sizes. Since for the complete graph, the complexities grow exponentially, it is not possible to simulate the process even for small population sizes. Moreover, for small population sizes the 2-Hierarchy and the Hierarchy coincide. Hence we present simulation results for the empty graph, the binary tree, and the 2-Hierarchy.

1. *Fixed ℓ and varying n.* We consider $\ell = 10$, and vary population sizes from 10 to 1 000. For each population size and graph, we run 10 000 trials, and then take the average of the complexity measures. Our results are shown in Figure 4(a,d). We observe that 2-Hierarchy significantly improves over the empty graph in terms of rounds complexity.

2. *Fixed n and varying ℓ.* In Figure 4(b,c,e,f), we present the rounds complexity for fixed n and varying ℓ from 2 to 100. We use two different values of n: 30 and 100. We observe that even for $n = 30$ the 2-Hierarchy is better than the empty graph. Thus, even for small population the 2-Hierarchy graph is better than the empty graph.

Furthermore, in SI Section 5.3 we present simulation results for randomly generated population structures. Random graphs do not improve the complexity measures compared to the empty graph. In SI Section 5.4 we show the full distribution of the number of rounds to fixation, comparing empty graph, the 2-Hierarchy, and the Tree graph. Therein we also present analogous simulations for the case of non-symmetric overlaps among languages.

4 Further Directions

There are many possible directions for further research. Here we list those related to other types of learners and models of learning (see SI Section 6 for more suggestions):

One direction is to consider other types of learners, presumably with intermediate capabilities as compared to memoryless (p, q)-learners and powerful batch learners. Another direction is to consider populations comprising learners of different types.

Yet another direction is to extend the model by defining a notion of similarity among the languages in the search space of the learners. The potential implications of such a generalization are two-fold: First, one could consider learners who, when updating their hypothesis, preferably update to a language similar either to their current language or to the language of the speaker [38]. Second, instead of insisting that the learners converge to (exactly) the teacher's language, one could ask for the time to convergence to a language sufficiently similar to that of the teacher.

5 Discussion

A group of individuals, learning language from a teacher or from their linguistic environment, instantiate a novel evolutionary process. The learners formulate hypotheses, which get dismissed (or modified) if sentences are received that cannot be parsed. In a sense, the linguistic environment selects the correct grammar in an iterated, population based process over time. While the wrong grammars become extinct eventually, the correct grammar proliferates by eliciting copies of itself in other learners.

In the classical setting, the theory of learning by inductive inference considers a teacher and a learner. But here we have considered a group of learners. A new twist arises naturally: the learners not only listen to the teacher (or the environment) but also to each other. Communication between learners can be problematic, because a learner already holding the correct hypothesis can be thrown off by listening to another learner who entertains an incorrect hypothesis. We show that certain population structures increase the complexity of the overall learning task, while others reduce it. Hierarchical structures, which consist of layers of learners where each layer listens to the layer above, can be extremely efficient. Such structures might help in other types of structured cultural transmission.

In evolutionary graph theory, a population structure is represented by a graph, where each node is a type of an individual (such as either wild type or mutant), and the underlying evolutionary stochastic process in essence picks edges to update the type of individuals (for example in Moran process, an individual reproduces and then an edge is chosen for replacing one of its neighbours). In our scenario, each language hypothesis defines a type of the node of the graph and a stochastic process updates the language hypotheses. In evolutionary graph theory, fixation time represents the time till the population is homogeneous, which is precisely what we study as rounds complexity.

The process of learning language is akin to the endeavour of the scientific progress. Here nature is the teacher, natural laws are the grammatical rules, and scientists are the learners. Scientists listen to evidence from nature and also listen to each other. Sometimes scientists hold wrong hypothesis and thereby confuse others. The communication of scientific knowledge has some hierarchical structures: from scientists to science teachers to students. Our results suggest that communication between individuals, although potentially confounding, can increase the overall efficiency of the process.

6 Methods

In this section we briefly describe our key methods to establish both upper and lower bounds for the various complexity measures.

Construction of graphs. The first key step in achieving our results is the construction of the graphs. Intuitively the tree graph presents an approach of learning in different levels with distributed responsibility for teaching. The 2-Hierarchy graph is based on the intuition that we first make a small group of people learn, and then they become teachers as well. The Hierarchy extends the idea of 2-Hierarchy iteratively.

Bounds for measures. Our upper bound for (p, q)-learners on the tree graph is based on an analysis of the process and uses Chernoff bound [39]. For the 2-Hierarchy and Hierarchy graphs, the principle is that once a group learns the language of the teacher, it teaches the next group. For every group of learners, we define its *phase* as lasting from the moment everyone in all the previous groups speaks the right language until everyone in that group also speaks the right language. We establish the number of rounds each phase takes and obtain the desired result by summing over all the groups. For batch learners, we proceed similarly. See SI for details.

Lower bound. The most interesting lower bound we establish is on the communication complexity, as we derive all other lower bounds from it. We actually show that for (p, q)-learners, no graph can achieve a communication complexity better than $c \cdot n \log n$, for some constant $c > 0$. For the result we use a coupling argument [40] to compare an arbitrary graph with the empty graph, and use Markov's inequality [39].

Author contributions

R.I.-J., J.T., K.C., and M.A.N. designed research, performed research, and wrote the paper.

Data Accessibility

Data and scripts for plotting figures have been uploaded as part of the electronic supplementary material.

Funding Statement

A.P., J.T. and K.C. acknowledge support from ERC Start grant no. (279307: Graph Games), Austrian Science Fund (FWF) grant no. P23499-N23 and S11407-N23 (RiSE). M.A.N. acknowledges support from Office of Naval Research grant N00014-16-1-2914 and from the John Templeton Foundation. The Program for Evolutionary Dynamics is supported in part by a gift from B. Wu and E. Larson.

Competing interests

We have no competing interests.

References

[1] Lightfoot D. The development of language: Acquisition, change, and evolution. Wiley-Blackwell; 1999.

[2] Wexler K, Culicover P. Formal principles of language acquisition. MIT Press; 1980.

[3] Smith JM. Evolution and the Theory of Games. Cambridge university press; 1982.

[4] Komarova NL, Niyogi P, Nowak MA. The evolutionary dynamics of grammar acquisition. Journal of theoretical biology. 2001;209(1):43–59.

[5] Nowak MA. Evolutionary dynamics. Harvard University Press; 2006.

[6] Niyogi P. The computational nature of language learning and evolution. MIT press Cambridge, MA:; 2006.

[7] Nowak MA, Komarova NL, Niyogi P. Computational and evolutionary aspects of language. Nature. 2002;417(6889):611–617.

[8] Chomsky N, DiNozzi R. Language and mind. Harcourt Brace Jovanovich New York; 1972.

[9] Jain S, Osherson D, Royer JS, Sharma A. Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change). 2nd ed. The MIT Press; 1999.

[10] Vapnik VN, Vapnik V. Statistical learning theory. vol. 1. Wiley New York; 1998.

[11] Gold EM. Language identification in the limit. Information and control. 1967;10(5):447–474.

[12] Osherson DN, Stob M, Weinstein S. Systems that learn: An introduction to learning theory for cognitive and computer scientists. The MIT Press; 1986.

[13] Pinker S. Formal models of language learning. Cognition. 1979;7(3):217–283.

[14] Niyogi P, Berwick RC. A language learning model for finite parameter spaces. Cognition. 1996;61(1):161–193.

[15] Osherson DN, Stob M, Weinstein S. Learning theory and natural language. Cognition. 1984;17(1):1–28.

[16] Case J, Moelius Iii SE. Optimal language learning. In: Algorithmic Learning Theory. Springer; 2008. p. 419–433.

[17] Heinz J, Kasprzik A, Kötzing T. Learning in the limit with lattice-structured hypothesis spaces. Theoretical Computer Science. 2012;457:111–127.

[18] Chomsky N. Principles and parameters in syntactic theory. Explanation in linguistics: The logical problem of language acquisition. 1981;32:75.

[19] Yang CD. Knowledge and learning in natural language. Oxford University Press on Demand; 2002.

[20] De la Higuera C. Grammatical inference: learning automata and grammars. Cambridge University Press; 2010.

[21] Heinz J, Sempere JM. Topics in grammatical inference. Springer; 2016.

[22] Zeugmann T. From learning in the limit to stochastic finite learning. Theoretical Computer Science. 2006;364(1):77–97.

[23] Nowak MA, Komarova NL, Niyogi P. Evolution of universal grammar. Science. 2001;291(5501):114–118.

[24] Komarova N, Rivin I. Mathematics of learning. arXiv preprint math/0105235. 2001;.

[25] Christiansen MH, Dale RA, Ellefson MR, Conway CM. The role of sequential learning in language evolution: Computational and experimental studies. In: Simulating the evolution of language. Springer; 2002. p. 165–187.

[26] Komarova NL, Nowak MA. Language dynamics in finite populations. Journal of Theoretical Biology. 2003;221(3):445–457.

[27] Lee Y, Stabler TCCEP, Taylor CE. The role of population structure in language evolution. language. 2005;22(23):24–25.

[28] Nowak MA, Krakauer DC. The evolution of language. Proceedings of the National Academy of Sciences. 1999;96(14):8028–8033.

[29] Stabler EP. Mathematics of language learning. Histoire Épistémologie Langage. 2009;31(1):127–145.

[30] Niyogi P, Berwick RC. Evolutionary consequences of language learning. Linguistics and Philosophy. 1997;20(6):697–719.

[31] Niyogi P, Berwick RC. The proper treatment of language acquisition and change in a population setting. Proceedings of the National Academy of Sciences. 2009;106(25):10124–10129.

[32] Kirby S. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. Evolutionary Computation, IEEE Transactions on. 2001;5(2):102–110.

[33] Clark R, Roberts I. A computational model of language learnability and language change. Linguistic Inquiry. 1993;24(2):299–345.

[34] Cormen TH. Introduction to algorithms. MIT press; 2009.

[35] Attiya H, Welch J. Distributed computing: fundamentals, simulations, and advanced topics. vol. 19. John Wiley & Sons; 2004.

[36] Lynch NA. Distributed algorithms. Morgan Kaufmann; 1996.

[37] Tanenbaum AS, Wetherall D. Computer networks. Prentice hall; 1996.

[38] Bryden J, Wright SP, Jansen VA. How humans transmit language: horizontal transmission matches word frequencies among peers on Twitter. Journal of The Royal Society Interface. 2018;15(139):20170738.

[39] Mitzenmacher M, Upfal E. Probability and computing: Randomized algorithms and probabilistic analysis. Cambridge University Press; 2005.

[40] Lindvall T. Lectures on the coupling method. Courier Corporation; 2002.

Figure and table captions

Figure 1. A teacher and a group of learners. The teacher is represented as a square and learners as circles. Individuals whose hypothesis is the teacher's language L_1 are shown in red, others in blue (teacher is always red). Possible communications are indicated by edges. When an edge is selected for the communication event, it is shown in bold. **(a)** An illustration of (p, q)-learning. In one step of the learning process, we select an edge (indicated in bold) and then the listener of that edge updates their language hypothesis. (i) Learner X listens to the teacher and switches to the teacher's language with probability p. (ii) Learner Y already has the same language as the teacher, but due to listening to a learner X who speaks a 'wrong' language, Y switches with probability $1-q$ to a (possibly different) wrong language. **(b)** An illustration of one possible run of a single round as described in the paragraph Example. Population structure consists of a teacher, Alice, and Bob. There are two non-overlapping languages L_1, L_2. When a learner hears a sentence they don't understand, they switch their hypothesis to the other language with probability 80 % (and keep it otherwise). We picked the edges in order $B \to T$, $B \to A$, $A \to T$. In the second step, B switched from correct L_1 to incorrect L_2.

Figure 2. Simulations for small graphs. (a) Four distinct structures of the class room, each with one teacher and four learners. Note that Graph A is the 'empty graph' because there are no communications between the learners. **(b)** Simulation results for these four graphs showing the average number of rounds that are needed for all learners to converge to the correct language versus the number of languages ℓ in the search space. Here we consider (p, q)-learners with $p = q = 1/\ell$. Each point is an average over 100 000 trials. In each round, the communication happens along each edge once, in random order. Graphs B and C are much worse than the empty graph, A, but graph

D is faster. This simple example shows that communication between learners can both accelerate and decelerate the process.

Figure 3. Different population structures of language learning. The teacher is shown in red and the learners in blue. **(a)** The empty graph represents the case where learners only listen to the teacher and do not communicate with each other. **(b)** The opposite extreme is the complete graph where all possible communications between learners are realized. **(c)** In the tree graph with branching factor $k = 2$, the teacher speaks to two learners, who each speak to two learners and so on. **(d, e)** The 2-Layered Hierarchy and the k-Layered Hierarchy consist of layers such that each learner from a given layer listens to all individuals from the previous layer. In the special case of Exponentially growing Layered Hierarchies (2-Hierarchy and Hierarchy), each layer is exponentially bigger than the previous one.

Figure 4. Numerical simulation results. The colours represent different graph families: Blue: Empty graph; Orange: 2-Hierarchy; Green: Tree graph. The empty graphs is shown in bold since it is the baseline comparison. First, we consider memoryless learners with helpful teacher, that is $p = 2/\ell$, $q = 1/\ell$ **(a)** Rounds complexity against the population size n, for fixed number of languages $\ell = 10$. For empty graph the dependency on n is logarithmic, for tree graph it is also logarithmic but worse by a constant factor, and for the 2-Hierarchy graph it is asymptotically better (namely doubly logarithmic). **(b), (c),** Rounds complexity against the number of languages ℓ, for fixed population size $n = 30$ and $n = 100$. The 2-Hierarchy beats the empty graph in both cases. Since the dependency on ℓ in all cases is linear, any value of ℓ would yield analogous outcome in (a). **(d), (e), (f)** Similar plots for batch learners under symmetric language overlap $q = 0.1$. **(d)** Rounds complexity against the population size n, for fixed number of languages $\ell = 10$. As in (a), for the empty graph the dependency is logarithmic whereas for the 2-Hierarchy it is asymptotically better. However, for tree graph the dependency is linear in n. **(e), (f)** This time the dependency on ℓ is logarithmic in all cases (batch learners are more powerful than memoryless learners). All the values shown are averages over 10 000 trials.

Table 1. Complexity bounds for language learning. The tables show the various complexity measures for different graphs as function of population size, n, and expected time to teach one learner in a single teacher single learner model, T. The first table refers to (p, q)-learners, the second table refers to batch learners under symmetric language overlap. Rounds complexity denotes the

average number of rounds until all learners hold the correct grammar. Communication complexity denotes the average number of communications until this state is reached and bottleneck complexity denotes the average maximum number of communications produced from a single person. There exist constants c_1, c_2, c_3, c_4 such that the complexity measures are lower bounded by the expressions. Except for batch learners on tree graphs, all bounds are tight up to a constant, which means there exist positive constants for which the corresponding expressions are upper bounds. The expression $\log^\star n$ denotes the iterated logarithm of n (see text).

Supplementary Information: Language acquisition with communication between learners

Rasmus Ibsen-Jensen*†, Josef Tkadlec*†, Krishnendu Chatterjee†, Martin A. Nowak‡

Submitted to Journal of the Royal Society Interface

Contents

*These authors contributed equally to this work
†IST Austria, Klosterneuburg, A-3400, Austria
‡Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, MA 02138, USA

1 Overview: Model and Results

In this section, we present our model for learning and then the theoretical results for our learning model. The results, presented later in the section, subsume the results described in the main article.

1.1 Model

We introduce our model of learning below.

Language learning: single learner. In the traditional setting, there is a set L of ℓ languages, a teacher who speaks a language from L, and a *single* learner who is trying to infer the teacher's language. At the beginning of the learning process, the learner has a hypothesis $L_i \in L$ as for what the teacher's language is. In each step, the teacher speaks a sample sentence from his language. Then the learner updates her hypothesis according to some learning algorithm of hers. The process stops when the learner's hypothesis matches the teacher's language.

The efficiency of the learning process can be measured by the expected number T of steps required by the learner to infer the teacher's language. Clearly, T depends on many factors, including the properties of the set L of languages; the learning algorithm employed by the learner; the way the teacher picks the sentences; or the presence or absence of noise.

Structured language learning: n learners. In this work, we consider the problem of language learning for a set of n learners who communicate with each other in a structured way. The structured communication is captured by a directed graph among the individuals. The vertices of the graph are the learners and also the teacher, and the edges capture possible communication. If there is an edge $e = (i, j)$ from vertex i to vertex j, then we say i is the *listener* of the edge e and j is its *speaker*. A *communication* along an edge is a sentence from the speaker of that edge to the listener, that is, the use of an edge of the graph as a sentence communication.

As in the scenario with single learner, there is a set L of ℓ languages and a teacher who speaks a language from L. At the beginning, all learners have their own hypotheses $L_i \in L$. The process proceeds in *rounds*, where each round consists of two stages:

1. First, all the edges of the graph are sorted in a random order.

2. Second, these edges are chosen for *communication* in the selected order. That is, the speaker of the edge speaks a sentence from his language and the listener of that edge updates her hypothesis according to her learning algorithm.

The process stops when the hypothesis of each learner matches the language of the teacher.

Main question. The main scientific question is: Which communication structures speed up the learning process? To answer this question, we do the following:

1. We consider several complexity measures (rounds complexity, communication complexity, bottleneck complexity) that can be used to evaluate the efficiency of the process for n learners.
2. We consider several communication structures (Empty graph, Tree graph, Hierarchy graphs) and evaluate them using the measures described above.
3. We consider two different types of learning algorithms (weak memoryless (p, q)-learners, and powerful batch learners) that cover the opposite ends of the learning capabilities spectrum.

Given a complexity measure, a communication structure, and a learning algorithm employed by the learners, we express the complexity measure in terms of the population size n and the expected time T to convergence for a single learner. We are mostly interested in dependence on n.

Complexity measures. We consider the following complexity measures to determine the efficiency of the process.

1. **Rounds complexity.** The rounds complexity is the expected number of rounds till the process converges (i.e., the average number of rounds till everyone has the same chosen language). This represents the total time, if each round takes constant time.
2. **Communication complexity.** The communication complexity is the expected number of communications till the process converges (i.e., the expected number of edge usages in the graph). This represents the total number of communications occurring.
3. **Bottleneck complexity.** The bottleneck complexity is the expectation of the maximum number of communications that needs to be done by a single individual (the teacher or one of the learners). This represents the bottleneck of the learning process.

Communication Structures. We consider the following classes of graphs. The *empty graph* serves as a natural benchmark. See Figure 3 from the main text for illustrations.

1. *Empty graph.* An *empty graph* E_n with n learners is the graph on $n+1$ vertices in which each learner listens to the teacher and no learner listens to any other learner. The Empty graph corresponds to no communication among learners.
2. *Complete graph.* A *complete graph* K_n with n learners is the graph on $n+1$ vertices in which each learner listens to the teacher and to every other learner.
3. *Tree graph.* A *tree graph* T_n with n learners is a complete binary tree on $n+1$ vertices rooted at the teacher. In more detail, the vertices are organized in layers with the first layer containing only the teacher and every other layer (except, possibly, the last one) containing twice as many vertices than the previous one. Each learner listens to exactly one individual from the previous layer and each individual speaks to at most two listeners. The last layer is "filled in from the left".
4. *Layered Hierarchy graphs.* Intuitively, in Layered Hierarchy graphs the learners are arranged in groups of gradually increasing sizes. Then we include all edges from the first group to the teacher and in general all edges from the next group to the previous one. By choosing the group sizes accordingly, we obtain several distinct notions of Layered Hierarchy graphs. See Section 3.6 for particular cases (2-Hierarchy, Hierarchy).

Learning algorithms. We consider two types of learners: weak memoryless (p, q)-learners and powerful batch learners.

1. (p, q)**-learners.** A (p, q)-learner is described using two positive parameters p, q such that $p + q < 1$. When receiving a sample sentence from a speaker, a (p, q)-learner does as follows:
   (a) If the learner has the same language as the speaker, then nothing changes.
   (b) Otherwise, the learner updates her language as follows:
   - with probability p, the learner changes to the language of the speaker;
   - with probability q, the learner does not change her language; and
   - otherwise, the learner switches to one of the other languages uniformly at random (that is, the learner switches to each of the remaining languages with probability $(1 - p - q)/(\ell - 2)$).
2. **Batch learners.** A batch learner always hypothesizes the language that is most consistent with all the sentences received so far. That is, having heard k sentences, let $c_1, c_2, \ldots, c_\ell$ be the number of sentences consistent with languages $L_1, L_2, \ldots, L_\ell$, respectively, and assume that the maximum of $c_1, c_2, \ldots, c_\ell$ is $c_{\max}$. Then a batch learner updates her hypothesis to $L_{\max}$.

1.2 Theoretical results

We present theoretical results related to the effect of the graph structure on the complexity measures. First, we focus on (p, q)-learners for fixed p, q. Second, we consider batch learners under symmetrically overlapping languages.

Notation. We will use the following standard notations: we use $O(\cdot)$ to denote asymptotic upper bounds, $\Omega(\cdot)$ to denote asymptotic lower bounds, and when the asymptotic upper and lower bounds match we denote it by $\Theta(\cdot)$. In the bounds below we only show the dependency on n and omit the dependency on T.

1.2.1 (p, q)-learners

To start, as a baseline comparison we first consider the classical model where there is no communication between the learners.

Classroom teaching: empty graph. We establish the following bounds (see Theorem 1): (a) The rounds complexity is $\Theta(\log n)$; (b) the communication complexity is $\Theta(n \log n)$; and (c) the bottleneck complexity is $\Theta(n \log n)$ (which is the communication complexity of the teacher). Moreover, we show that for communication complexity the empty graph is optimal (i.e., no graph structure can achieve communication complexity better than the empty graph) (see Theorem 2). The results for empty graph are summarized in the first row of Table 1.

Complete graph. For the complete graph of communication between the learners we establish that all the complexity measures are at least exponential. Hence the complete graph is dramatically worse as compared to the empty graph. The results for complete graph are summarized in the second row of Table 1.

Intuitive description. We present the intuitive reason for the result on the complete graph.

We observe that the process is biased as follows: At any given step of round, we can divide the learners into groups A (also containing T) and B of those who speak the teacher's language and those who do not, respectively. Consider the case when B is much smaller than A (i.e., in terms of

	Rounds Compl.	Communication Compl.	Bottleneck Compl.
Empty graph	$\Theta(T \cdot \log n)$	$\Theta(T \cdot n \log n)$	$\Theta(T \cdot n \log n)$
Complete graph	$\Omega(c^n)$	$\Omega(c^n)$	$\Omega(c^n)$
Tree	$\Theta(T \cdot \log n)$	$\Theta(T \cdot n \log n)$	$\Theta(T \cdot \log n)$
2-Hierarchy	$\Theta(T \cdot \log \log n)$	$\Theta(T \cdot n \log n)$	$\Theta(T \cdot n \log \log n)$
Hierarchy	$\Theta(T \cdot \log^\star n)$	$\Theta(T \cdot n \log n)$	$\Theta(T \cdot n \log^\star n)$

Table 1: **Complexity measures for (p, q)-learners in terms of T and n.** The various complexity measures for different graphs as a function of population size n, and the time to convergence for a single learner T. Here p, q are considered fixed.

set sizes we are close to completion). If we pick an edge from A to A, then nothing happens. If we pick an edge from B to B, then, in case they do not speak the same language, we "gain" the learner with chance $(1 - p - q)/(\ell - 2)$. If we pick an edge from some $b \in B$ to some $a \in A$, we "gain" b with chance p. Finally, if we pick an edge from $a \in A$ to $b \in B$, we "lose" a with chance $1 - q$. The probability to pick (1) an edge from A to B is nearly equal to picking (2) an edge from B to B or from B to A. More precisely, the chance of picking an edge from B to B is small (since B is small) and the probability of picking an edge from B to the teacher is also small. The probability of picking each other edge from A to B is equal to the probability of picking the reversed edge, that goes from B to A. But if we pick an edge in (1) we "lose" one learner with probability $1 - q$ and if we pick an edge in (2) we "gain" one leaner only with probability p. Since $1 - q > p$, we lose with large probability and gain with only small probability. This bias ensures that the process takes exponentially long to converge.

Tree graph. For the binary tree graph which has at most $\log n$ levels we establish the following results (see Theorem 5): (a) The rounds complexity is $\Theta(\log n)$; (b) the communication complexity is $\Theta(n \log n)$; and (c) the bottleneck complexity is $\Theta(\log n)$. We observe that the tree graph achieves the same asymptotic rounds and communication complexity as the empty graph. Moreover, it intuitively distributes the teaching responsibility from single teacher to the learners, thus reducing the bottleneck complexity. The results for tree graph are summarized in the fourth row of Table 1.

Layered Hierarchy graphs. Our most interesting results are related to various notions of Layered Hierarchy graphs which provide a significant improvement over the empty graph among learners. We start with the simpler case of a 2-Hierarchy and then the Hierarchy graph. We establish the following results (see Theorem 6 and Theorem 7).

1. **2-Hierarchy.** For 2-Hierarchy graphs we show: (a) The rounds complexity is $\Theta(\log \log n)$; (b) the communication complexity is $\Theta(n \log n)$; and (c) the bottleneck complexity is $\Theta(n \log \log n)$. Hence the 2-Hierarchy achieves the same asymptotic communication complexity as the empty graph, however, it improves the bottleneck complexity of empty graph from $n \log n$ to $n \log \log n$, and improves the rounds complexity exponentially (hence significantly) from $\log n$ to $\log \log n$.
2. **Hierarchy.** For Hierarchy (which generalizes 2-Hierarchy) we show: (a) The rounds complexity is $\Theta(\log^\star n)$; (b) the communication complexity is $\Theta(n \log n)$; and (c) the bottleneck complexity is $\Theta(n \log^\star n)$. Recall that $\log^\star n$ is the iterated logarithm of n (usually read "log star"), and can be described recursively as

$$\log^\star(n) = \begin{cases} 0 & n \leq 1 \\ 1 + \log^\star(\log n) & n > 1 \end{cases}$$

	Rounds Compl.	Communication Compl.	Bottleneck Compl.
Empty graph	$\Theta(T + \log n)$	$\Theta(T \cdot n + n \log n)$	$\Theta(T \cdot n + n \log n)$
Complete graph	∞	∞	∞
Tree	$\Omega(T \cdot n)$	$\Omega(T \cdot n^2)$	$\Omega(T \cdot n)$
2-Hierarchy	$\Theta(T + \log \log n)$	$\Theta(T \cdot \frac{n \log n}{\log \log n} + n \log n)$	$\Theta(T \cdot n + n \log \log n)$

Table 2: **Complexity measures for batch learners with symmetric overlap in terms of T and n.** The various complexity measures for different graphs as a function of population size n, and the time to convergence for a single learner T. Here the overlap q is considered fixed.

The values of $\log^\star(n)$ for different values of n are as follows:

$$\log^\star n = \begin{cases} 0 & n = 1 \\ 1 & 1 < n \leq 2 \\ 2 & 2 < n \leq 4 \\ 3 & 4 < n \leq 16 \\ 4 & 16 < n \leq 65536 \\ 5 & 65536 < n \leq 2^{65536} \\ \dots & \end{cases}.$$

Note that for $n = 2^{65536}$ we have $\log^\star n = 5$ and thus $\log^\star n$ is effectively constant for all practical purposes. We note that Hierarchy provides dramatic improvements over the empty graph: the communication complexity for the Hierarchy asymptotically matches with the empty graph, whereas the rounds complexity is improved from $\log n$ to almost a constant (i.e., $\log^\star n$) and the bottleneck complexity is improved from $n \log n$ to effectively n. Thus the Hierarchy graph is much better as compared to the empty graph. The results for Hierarchy graphs are summarized in the final rows of Table 1.

Remark on asymptotic complexity. Note that as mentioned in the results for empty graph (see Theorem 2), the empty graph is optimal for communication complexity. Hence no graph can be better than the empty graph for communication complexity. Our results show that the tree graph and the Hierarchy graphs achieve the same asymptotic communication complexity, however, the associated constants are worse than for the empty graph. In other words, for example, 2-Hierarchy is worse by a constant factor than empty graph for communication complexity, but asymptotically better in terms of rounds and bottleneck complexity.

1.2.2 Batch learners

For batch learners under symetrically overlapping languages q we establish the following bounds.

Empty graph. The rounds complexity is $\Theta(\log n)$, the communication complexity is $\Theta(n \log n)$, and the bottleneck complexity is $\Theta(n \log n)$ (see Theorem 7).

Complete graph. The complexity measures are undefined because the process doesn't terminate with probability 1 (see Lemma 6).

Tree graph. The rounds complexity is $\Omega(n)$, the communication complexity is $\Omega(n^2)$, and the bottleneck complexity is $\Omega(n)$ (see Theorem 8).

2-Hierarchy graph. As with the (p, q)-learners, the rounds complexity is $\Theta(\log \log n)$, the communication complexity is $\Theta(n \log n)$, and the bottleneck complexity is $\Theta(n \log \log n)$ (see Theorem 9).

2 Formal Model

In this section we formally define the stochastic learning processes that model structured learning, three complexity measures for such processes, namely rounds complexity, communication complexity, and bottleneck complexity, and the graphs we consider.

2.1 Communication graph and its labelling

The individuals are represented as vertices of a directed graph with a directed edge (u, v) meaning that individual u can listen to what individual v says. The finite set $L = \{L_1, \ldots, L_\ell\}$ of ℓ languages (grammars) is represented by numbers $1, \ldots, \ell$ used as labels for these vertices.

Communication Graph. Formally, a *communication graph* $G = (V, E, T)$ is a directed graph with a set V of vertices, set $E \subseteq V \times V$ of directed edges and a distinguished vertex T. The distinguished vertex is called the *teacher*, the other vertices are called *learners*, all of them are called *individuals*. A vertex v is called *active* if it has at least one incoming edge (i.e. there exists $u \in V$ such that $(u, v) \in E$). The *indegree* $\mathrm{In}(v)$ of a vertex v is the number of its incoming edges, i.e. $\mathrm{In}(v) = |\{u \mid (u, v) \in E\}|$. Given an edge $(u, v) \in E$, we refer to u as *listener* and to v as *speaker* of that edge.

Labelling. Given a communication graph $G(V, E, T)$ and a number $\ell \in \mathbb{N}$ of languages, a *labelling* l of G by labels $\{1, 2, \ldots, \ell\}$ is any function l that assigns a label to each vertex (i.e. $l \colon V \to \{1, 2, \ldots, \ell\}$). A vertex is called *convinced* if its label is the same as teacher's label and *bad* otherwise. A *labelled communication graph* is a 5-tuple $G(V, E, T, \ell, l)$. Informally, a labeled communication graph is a state capturing what are the current hypotheses of the respective learners as for the teacher's language in the current time step. The original labelling is given by $l(T) = 1$ for the teacher and $l(v) = \ell$ for all the learners.

2.2 Learning algorithms

Now we describe how the labelling changes in time.

Using an edge. Given a labelled communication graph, the learning process is a stochastic process that proceeds in rounds. In each round, the edges of the graph are sorted in random order (uniformly at random and independently on the other rounds) and then they are *used* in that order. When the edge is used, its speaker produces a sample sentence from his language (label) and the listener receives the sentence and updates her hypothesis (label) according to her learning algorithm.

Learning algorithms: (p, q)-learning and batch learning. Here we describe how listener can update her language hypothesis.

1. *(p, q)-learning.* A (p, q)-learner is described using two positive parameters p, q such that $p + q < 1$. Assume an edge $(u, v) \in E$ is being used. If the speaker's current hypothesis is $l(v)$ and the learner's current hypothesis is $l(u)$, then the learner updates to new hypothesis $l'(u)$ as follows:
   (a) If $l(u) = l(v)$ then $l'(u) = l(u)$.

(b) If $l(u) \neq l(v)$ then $l'(u)$ is a probability distribution such that

$$\begin{aligned}\mathbb{P}[l'(u) = l(v)] &= p,\\ \mathbb{P}[l'(u) = l(u)] &= q,\\ \mathbb{P}[l'(u) = i] &= \frac{1-p-q}{\ell-2} \quad \text{for all } 1 \leq i \leq \ell \text{ different from } l(u),\, l(v).\end{aligned}$$

2. *Batch learning.* Having heard k sentences, a batch learner forms a hypothesis as follows: For each $i = 1, \ldots, \ell$ let c_i be the number of received sentences consistent with language L_i. Let $c_{\max}$ be the largest of these numbers (in case of a tie, take such number for which the index max is largest). Then a batch learner takes $L_{\max}$ to be her hypothesis.
   We will investigate batch learners in the case when the languages have symmetric overlap q. That is, whenever a speaker speaks a sentence from language L_i then the sentence is consistent with any other language L_j, $j \neq i$, with constant probability q independently on other languages and other sentences.

Remark on batch learning. Traditionally, a batch learner hypothesises a language consistent with all observed sentences. Note that in our setting, learners listen to speakers with different languages, and hence there might not be a single language that all sentences are consistent with. The batch learner thus updates to the language with maximum consistency. Also note that the notion of maximum consistency matches with the classical notion for a single teacher single learner scenario.

Stable labelling. We denote by $l^\star$ the *stable labelling*, i.e. the labelling assigning to every vertex the original label $l(T) = 1$ of the teacher. Intuitively, this corresponds to teacher having convinced all learners. The following lemma states that under some very reasonable assumptions on the structure of the underlying graph, the (p, q)-learners reach the stable labelling on average in finite number of rounds.

Lemma 1. *Suppose $G = (V, E, T, \ell, l)$ is a labelled communication graph such that*

- *for any vertex $v \neq T$ there exists a directed path from v to T (i.e. there exists a positive integer k and a sequence of vertices $v = u_0, u_1, \ldots, u_k = T$ such that $(u_{i-1}, u_i) \in E$ for all $i = 1, \ldots, k$); and*
- *vertex T has no outgoing edges (i.e. $(u, T) \notin E$ for all $u \in V$).*

Then for (p, q)-learners the expected number of rounds to reach a state with all vertices labelled by $l(T) = 1$ is finite.

Proof. By (ii), the teacher will never change his label. Assumption (i) implies that there exists at least one ordering of vertices (edges) such that if we use the vertices (edges) in that order and each switch happens to be a switch to language $l(T)$ then we reach the stable labelling in a single round. For (p, q)-learners, the chance that we reach the stable labelling in a single round is at least $p_0 = (1/|E|!) \cdot p^{|E|} > 0$, hence the expected number of rounds is finite and at most $1/p_0 < \infty$. □

2.3 Graphs

Here we define the classes of graphs we will analyze. Note that all these graphs satisfy the assumptions of Lemma 1. The *empty graph* serves as a natural benchmark. See Figure 3 from the main text for illustrations.

1. *Empty graph.* An *Empty graph* E_n with n learners is the graph on $n + 1$ vertices in which each learner listens to the teacher and no learner listens to any other learner. Formally, $V = \{T, v_1, \ldots, v_n\}$ and $E = \{(v_i, T) \mid i = 1, \ldots, n\}$.

2. *Complete graph.* A *complete graph* K_n with n learners is the graph on $n+1$ vertices in which each learner listens to the teacher and to every other learner. Formally, $V = \{T, v_1, \ldots, v_n\}$ and $E = \{(v_i, T) \mid i = 1, \ldots, n\} \cup \{(v_i, v_j) \mid i, j = 1, \ldots, n, i \neq j\}$.
3. *Tree graph.* A *tree graph* T_n with n learners is a binary tree on $n+1$ vertices rooted at the teacher. Formally, $V = \{T = v_0, v_1, \ldots, v_n\}$ and $E = \{(v_i, v_j) \mid i = 2j+1 \text{ or } i = 2j+2\}$.
4. *Layered Hierarchy graphs.* Intuitively, in Layered Hierarchy graphs the learners are arranged in groups of gradually increasing sizes. Then we include all edges from the first group to the teacher and in general all edges from the next group to the previous one. Formally, given a finite sequence of integers $s = (s_1, \ldots, s_k)$, a *Layered Hierarchy graph* determined by s is the following graph $\mathrm{H}(s) = \mathrm{H}(s_1, s_2, \ldots, s_k)$:
   - Vertices: For each $i = 1, \ldots, k$ define S_i as a set of s_i vertices. Also, set $S_0 = \{T\}$ to be the teacher. Hence we get $1 + \sum_{i=1}^{k} s_i$ vertices.
   - Edges: For every $i = 1, \ldots, k$, every $u \in S_i$ and every $v \in S_{i-1}$ include a directed edge $u \to v$.

   By choosing the sequence s accordingly, we obtain several distinct notions of Layered Hierarchy graphs. See Section 3.6 for particular cases (2-Hierarchy, Hierarchy).

2.4 Complexity measures

Next we define the complexity measures.

Random variable $R(G)$. Let $G(V, E, T)$ be a communication graph and $\ell \geq 2$ a number of languages. For definiteness, we set the initial labelling l_0 to $l_0(T) = 1$ and $l_0(v) = \ell$ for all $v \neq T$. As proved in Lemma 1, on all the graphs discussed, the (p, q)-learners reach the stable labelling in finite number of rounds with probability one. Let $R(G)$ be a random variable capturing how many rounds the process takes to reach the stable labelling $l^\star$ from the initial labelling l_0. Based on the random variable $R(G)$ we define the following complexity measures.

1. *Rounds complexity.* We define the $\mathrm{Rnd}(G)$ as the expected number of rounds the process takes, i.e. $\mathrm{Rnd}(G) = \mathbb{E}[R(G)]$.
2. *Communication complexity.* Denote by $m = |E|$ the number of edges in G. Then the communication complexity is $\mathrm{Comm}(G) = m \cdot \mathrm{Rnd}(G)$. Communication complexity is the expected number of edge usages until the process stops.
3. *Bottleneck complexity.* Denote by d the maximum indegree among the vertices in G. Then the bottleneck complexity is $\mathrm{Bot}(G) = d \cdot \mathrm{Rnd}(G)$. If we imagine that communicating along edge (u, v) incurs cost 1 to its speaker v then bottleneck complexity is the expected value of the largest total cost incurred by a vertex. When bottleneck complexity is large, some individuals are used heavily during the communication process which can be viewed as the process having a bottleneck.

For batch learners under symmetric language overlap q we define the complexity measures $\mathrm{Rnd}_b(G)$, $\mathrm{Comm}_b(G)$, $\mathrm{Bot}_b(G)$ analogously. The type of learner will always be understood from context.

2.5 Basic Mathematical Tools

Here we summarize the standard mathematical tools and results that we use in the proofs, together with the references to the literature.

Proposition 1 (Cauchy-Schwarz inequality)**.** *(See Proposition B.9 in [1].) Let $u = (u_1, \ldots, u_n)$, $v = (v_1, \ldots, v_n)$ be vectors of n real numbers. Then*

$$\left(\sum_{i=1}^{n} u_i^2\right) \cdot \left(\sum_{i=1}^{n} v_i^2\right) \geq \left(\sum_{i=1}^{n} u_i v_i\right)^2.$$

In particular, for any $a_1, \ldots, a_n \in \mathbb{R}^+$ we have

$$\left(\sum_{i=1}^{n} a_i\right) \cdot \left(\sum_{i=1}^{n} \frac{1}{a_i}\right) \geq n^2.$$

Proposition 2 (Union Bound)**.** *(See Proposition C.2 in [1].) Let A_i, $i = 1, \ldots, n$, be events. Then*

$$\mathbb{P}\left[\bigwedge_{i=1}^{n} \overline{A_i}\right] \geq 1 - \sum_{i=1}^{n} \mathbb{P}[A_i].$$

Proposition 3 (Markov's inequality)**.** *(See Chapter 3, Theorem 3.2 in [1].) Let X be nonnegative random variable with expectation $\mu = \mathbb{E}[X]$ and let $\lambda > 0$ be a real number. Then*

$$\mathbb{P}[X > \lambda \cdot \mu] \leq \frac{1}{\lambda}.$$

Proposition 4 (Chernoff Bound)**.** *(See Chapter 4.1, Theorem 4.2 in [1].) Let $X_1, \ldots, X_n$ be independent random $\{0,1\}$-variables such that for $i = 1, \ldots, n$, $\mathbb{P}[X_i = 1] = p_i$, where $0 < p_i < 1$. Then, for $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$, and any $\delta > 0$,*

$$\mathbb{P}[X < (1-\delta)\mu] < \exp\left(-\mu\delta^2/2\right).$$

Proposition 5 (One-dimensional random walk)**.** *(See Chapter 7.7, proof of Theorem 7.1 in [2].) Let $\mathcal{M}$ be a discrete-time Markov chain with $n+1$ states $s_0, \ldots, s_n$ and transition probabilities $p_i \colon s_{i-1} \to s_i$, $q_i \colon s_i \to s_{i-1}$ $(i = 1, \ldots, n)$. Then the expected number of time steps to reach state s_n from state s_0 is denoted by $t_{0,n}$ and given by*

$$t_{0,n} = \sum_{1 \leq i \leq j \leq n} \frac{1}{p_j} \prod_{k=i}^{j-1} \frac{q_k}{p_k}$$

Technical comments. For brevity we omit the floor and ceiling symbols in all the computations unless they are crucial for the argument. Symbol $\log(n)$ denotes the natural logarithm of n (i.e. the logarithm to the base $e = \sum_{i=0}^{\infty} 1/i! = 2.718\ldots$).

3 (p, q)-learning

In this section we analyze the efficiency of (p, q)-learners on various graph families.

Throughout this section, we consider (p, q)-learners for fixed p, q and we investigate how the efficiency of the learning process depends on n.

3.1 Single teacher, single learner

First, we consider the baseline setting of a single teacher and single learner ($n = 1$). We show that the expected number T of sample sentences required to convergence for a single (p, q)-learner, is $1/p$. Later on, we will use this result to express the efficiency of the learning process in the scenario with n-learners in terms of n and T.

Lemma 2. *Let p, q be fixed. Then $T = 1/p$.*

Proof. A (p, q)-learner has a fixed probability p of converging to the teacher's language in one step, hence the expected number of steps is $T = 1/p$. Indeed, after every single step, the learner either needs no more sample sentences (with probability p) or, since she is memoryless, needs, on average, T sample sentences (with probability $1 - p$). Thus $T = 1 + p \cdot 0 + (1 - p) \cdot T$ which reduces to $T = 1/p$. □

3.2 Empty graph

Next theorem pinpoints $\mathrm{Rnd}(E_n)$ up to a constant factor and hence attests the first line of Table 1. Note that our bounds are independent of q.

Theorem 1. *Let E_n be the Empty graph on $n + 1$ vertices. Then*

$$\left(1 - \frac{1}{e}\right) \cdot \log_{\frac{1}{1-p}} n \leq \mathrm{Rnd}(E_n) \leq \log_{\frac{1}{1-p}} n + \frac{1}{p}.$$

Since $\log(1/(1-x)) \sim x$ for $x \to 0$, for fixed small p the above expressions can be approximated by

$$\left(1 - \frac{1}{e}\right) (1/p - 1) \cdot \log n \leq \mathrm{Rnd}(E_n) \leq (1/p - 1) \cdot \log n + \frac{1}{p},$$

that is $\mathrm{Rnd}(E_n) = \frac{1}{p} \cdot \Theta(\log n) = \Theta(T \cdot \log n)$.

Proof. Informally, in each round the teacher on average convinces a constant fraction (about p-portion) of the learners who are not convinced yet, so we expect to be done in about $\frac{1}{p} \cdot \log n$ steps, where $\log n$ is a natural logarithm.

Formally, for $k = 0, 1, \ldots$ let w_k be a random variable equal to the number of learners who are not convinced after the teacher spoke k times. Then $\mathrm{Rnd}(E_n) = \sum_{k=0}^{\infty} \mathbb{P}[w_k > 0]$. Clearly we have $\mathbb{E}[w_0] = w_0 = n$ and setting $\alpha = 1 - p < 1$, by linearity of expectation we obtain $\mathbb{E}[w_{k+1}] = \mathbb{E}[w_k] \cdot \alpha$ which gives $\mathbb{E}[w_k] \leq n \cdot \alpha^k$.

For $N = \log_{1/\alpha}(n)$ we thus get $\mathbb{E}[w_N] \leq 1$ and from that point on, $\mathbb{E}[w_{N+i}] \leq \alpha^i$. By Markov's inequality (see Proposition 3) we have $\mathbb{P}[w_k > 0] = \mathbb{P}[w_k \geq 1] \leq \mathbb{E}[w_k]$ for every k, hence

$$\mathrm{Rnd}(E_n) = \sum_{i=0}^{\infty} \mathbb{P}[w_i > 0] \leq \sum_{i=0}^{N-1} 1 + \sum_{i=N}^{\infty} \mathbb{E}[w_i] \leq N + \sum_{i=0}^{\infty} \alpha^i = N + \frac{1}{1-\alpha} = \lceil \log_{1/(1-p)}(n) \rceil + \frac{1}{p}.$$

On the other hand, $\mathbb{P}[w_k > 0] = 1 - (1 - \alpha^k)^n$ is a decreasing function of k and for $N = \log_{1/\alpha}(n)$ it is $\mathbb{P}[w_N > 0] = 1 - (1 - 1/n)^n \to 1 - 1/e$, hence

$$\mathrm{Rnd}(E_n) = \sum_{i=0}^{\infty} \mathbb{P}[w_i > 0] \geq \sum_{i=0}^{N-1} (1 - 1/e) + \sum_{i=N}^{\infty} 0 \geq \log_{1/\alpha}(n) \cdot (1 - 1/e).$$

□

As an immediate corollary, we get that the expected number of rounds for an Empty graph is asymptotically logarithmic in the number of learners. Since E_n has n edges and only one active vertex with indegree n, we easily compute the other complexity measures.

Corollary 1. *For (p,q)-learners we have*

$$\mathrm{Rnd}(E_n) = \Theta(T \cdot \log n), \qquad \mathrm{Comm}(E_n) = \Theta(T \cdot n \log n), \qquad \mathrm{Bot}(E_n) = \Theta(T \cdot n \log n).$$

3.3 Lower bounds

In this section we obtain general lower bounds for all three complexity measures. Later on we exhibit particular graphs witnessing that all these bounds are asymptotically tight.

The following lemma is the core of the proof of the lower bound for communication complexity. As it works for more general selection processes it might be of independent interest so we state it separately.

Lemma 3. *Let G be a graph with $n+1$ vertices $\{T, v_1, \ldots, v_n\}$ and m edges. Let p, q be positive real numbers such that $p+q<1$ and consider the initial labelling $l_0(T)=1$, $l_0(v_i)=2$ for $i=1,\ldots,n$. Let X be an infinite sequence of edges of G. Consider a stochastic process that changes the labelling by sequentially using the edges from X and denote by x the expected number of edge usages before the process reaches the stable labelling $l^\star(v_i)=1$ for $i=1,\ldots,n$. Then*

$$x \geq \frac{1}{2}\,\mathrm{Comm}(E_n).$$

Proof. For a given vector $a=(q_1,\ldots,q_n)\in\mathbb{N}^n$ of integer "quotas", denote by $[a]$ the equivalence class of a in $\mathbb{N}^n$, where $a\sim b$ provided that b is a permutation of a. Let $t_a(G)$ be the expected number of edge usages from X after which vertex v_i was a listener q_i times, for each $i=1,\ldots,n$. Similarly, for the empty graph E_n with $n+1$ vertices $\{T,w_1,\ldots,w_n\}$ let $u_a(E_n)$ be the expected number of edge usages under (p,q)-learning after which vertex w_i was a listener q_i times. We will prove that

$$\frac{1}{|[a]|}\sum_{b\in[a]} t_b(G) \geq \frac{1}{2}\frac{1}{|[a]|}\sum_{b\in[a]} u_b(E_n)$$

for every $a\in\mathbb{N}^n$. By a straightforward coupling argument, this will imply that $x\geq\frac{1}{2}\,\mathrm{Comm}(E_n)$.

Let $q^\star=\max_{i=1,\ldots,n}\{q_i\}$ be the largest quota (if there are more of them, we choose one uniformly at random). Then for any $b\in[a]$ we have $u_b(E_n)=q^\star\cdot n$ because in each of the first $q^\star$ rounds we use each of the n edges once. Thus $\frac{1}{2}\frac{1}{|[a]|}\sum_{b\in[a]}u_b(E_n)=q^\star\cdot\frac{n}{2}$.

Now we focus on $t_b(G)$. In order to reach the quota $q^\star$ on a vertex v_i, we need to take the prefix of X containing $q^\star$ edges that have v_i as a listener. Note that in a prefix of length $k\cdot q^\star$, at most k vertices appear as listeners at least $q^\star$ times. Since each of the n vertices is equally likely to receive this largest quota, we have

$$\frac{1}{|[a]|}\sum_{b\in[a]} t_b(G) \geq \frac{1}{n}\left(q^\star+2q^\star+\cdots+nq^\star\right) = q^\star\cdot\frac{n+1}{2} \geq q^\star\cdot\frac{n}{2}$$

as desired. □

Theorem 2. *Let G be an arbitrary graph with $n+1$ vertices $\{T,v_1,\ldots,v_n\}$ and m edges. Then*

1. $\mathrm{Rnd}(G) \geq 1$,
2. $\mathrm{Comm}(G) \geq \mathrm{Comm}(E_n) = \Omega(n \log n)$,
3. $\mathrm{Bot}(G) = \Omega(\log n)$,

Proof. We treat the three claims separately

1. This is trivial – we always need at least one round so in expectation we also need at least one round.
2. Since the statement of Lemma 3 holds for any sequence X of edges, it also holds for any distribution over sequences of edges, so it implies $\mathrm{Comm}(G) \geq \frac{1}{2}\mathrm{Comm}(E_n)$. To get rid of the factor $1/2$ we proceed analogously and provide a stronger bound particular to (p, q)-learning. Using the same notation as in the Lemma 3 we have $\frac{1}{|[a]|}\sum_{b\in[a]} u_b(E_n) = q^\star \cdot n$. Now denote by $d_i = \mathrm{Out}(v_i, G)$ the outdegree of v_i in G. Reaching the quota $q^\star$ on a vertex with outdegree d takes $\lceil q^\star/d \rceil$ rounds. Since each of the n vertices is equally likely to receive this largest quota and each round uses $m = \sum_{i=1}^n d_i$ edges, we have

$$\frac{1}{|[a]|}\sum_{b\in[a]} u_b(G) \geq \frac{1}{n}\left(\sum_{i=1}^n d_i\right)\cdot\left(\sum_{i=1}^n \frac{q^\star}{d_i}\right) \geq \frac{q^\star}{n}\cdot n^2 = q^\star \cdot n,$$

   where in the last inequality we used a particular case of Cauchy-Schwarz Inequality (see Proposition 1).
3. This follows from (b). Denote by d the maximum indegree among the vertices in G, by n_0 the number of active vertices, and by m the number of edges in G. Since $d \cdot n_0 \geq m$, we infer $\mathrm{Bot}(G) = d \cdot \mathrm{Rnd}(G) \geq \mathrm{Comm}(G, \ell)/n_0 = \Omega(\log n)$.

□

3.4 Complete graph

Here we analyze the rounds complexity on the complete graph.

The analysis on a complete graph is complicated by the fact that within each round, each edge of the n^2 edges in the graph has to be chosen precisely once and keeping track of the labelling only at the end of a round is technically involved. For cleaner exposition, we avoid these difficulties by analyzing a slightly different process called RndEdge selection. Intuitively, under RndEdge we don't require that each edge is used precisely once in each round but instead we use the edges one by one, randomly, and independently. Therefore, while the same edge might be used multiple times within a single round, it is still used precisely once within each round on average. This adjusted version of Edge selection process allows us to track the labelling after each edge usage.

We show that in all realistic scenarios, the rounds complexity of the complete graph under RndEdge selection is exponential in n.

RndEdge selection. Formally, denote by m the number of edges of G_n (for $G_n = K_n$ we have $m = n(n-1) + n = n^2$). The RndEdge selection is a stochastic process that proceeds in rounds where each round consists of m steps. In a single step, given a labelling l, the process produces a labelling l' obtained by using an edge selected uniformly at random, independently of the other steps. Hence within a single round, we still use m edges but it's possible that some of them are used multiple times while some others are not used at all. The process starts with labelling l_0 given by $l_0(T) = 1$ for the teacher and $l_0(v) = \ell$ for all $v \neq T$ and ends once $l_0(v) = 1$ for all v. Denote by $\mathrm{Rnd}_{re}(G)$ the expected number of rounds the RndEdge selection process takes on graph G, and define the measures $\mathrm{Comm}_{re}(G)$, $\mathrm{Bot}_{re}(G)$ accordingly.

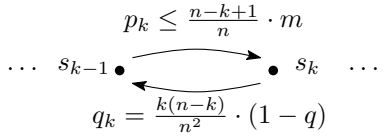

Figure SI.1: A part of the Markov chain associated with complete graph K_n under RndEdge selection. Self-loops are not depicted.

Theorem 3. *Let K_n be the complete graph with $n+1$ nodes. Assume that all p, q, $(1-p-q)/(\ell-2)$ are less than $1/4$. Then*

$$\mathrm{Rnd}_{re}(K_n) = \Omega(1.1^n)$$

Proof. Intuitively, at any time point in the process, it's more likely to lose a learner who is already convinced than to convince a learner who speaks a wrong language. This bias ensures that convincing all learners takes at least exponential time.

Formally, set $m = \max\{p, (1-p-q)/(\ell-2)\}$. By assumption, $m \le 1/4$.

We keep track of how many learners speak the teacher's language. Consider the Markov chain M with states $s_0, \dots, s_n$ corresponding to $0, \dots, n$ learners speaking the right language (see Figure SI.1 for illustration). The transition probability from s_{k-1} to s_k depends on how many learners speak various "wrong" languages but it is always at most $p_k \le \frac{n-k+1}{n} \cdot m$, since the listener of the selected edge originally has to speak the wrong language and then they have to switch to the right one. On the other hand, the transition probability from s_k to s_{k-1} is precisely $q_k = \frac{k(n-k)}{n^2} \cdot (1-q)$.

Since the considered Markov chain is one-dimensional, the expected number of steps to reach the absorbing state from state s_0 equals (see Proposition 5)

$$\mathrm{Comm}_{re}(K_n) = \sum_{1 \le i \le j \le n} \frac{1}{p_j} \prod_{k=i}^{j-1} \frac{q_k}{p_k} \ge \frac{1}{p_n} \cdot \frac{q_1}{p_1} \dots \frac{q_{n-1}}{p_{n-1}}$$

where we used one summand as a lower bound for the whole sum.

Observe that $\frac{q_k}{p_k} \ge \frac{n-k}{n-k+1} \cdot \frac{k}{n} \cdot \frac{1-q}{m}$. Plugging this in, we get

$$\mathrm{Comm}_{re}(K_n) \ge \frac{1}{p_n} \cdot \frac{1}{n} \cdot \frac{n!}{n^n} \cdot \left(\frac{1-q}{m}\right)^{n-1} > \frac{m}{(1-q)np_n} \left(\frac{1-q}{e \cdot m}\right)^n = \frac{1}{1-q} \left(\frac{1-q}{e \cdot m}\right)^n,$$

where we used $\prod_{k=1}^{n-1} \frac{n-k}{n-k+1} = \frac{1}{n}$, then $n! > (n/e)^n$, and then $p_n = m/n$.

By assumption, $(1-q)/(em) > 3/e > 1.1$, hence $\mathrm{Rnd}_{re}(K_n) = \mathrm{Comm}_{re}(K_n)/n^2 = \Omega(1.1^n)$. □

Corollary 2. *Assume that all p, q, $(1-p-q)/(\ell-2)$ are less than $1/4$. Then for (p,q)-learners we have*

$$\mathrm{Rnd}_{re}(K_n) = \Omega(1.1^n), \qquad \mathrm{Comm}_{re}(K_n) = \Omega(1.1^n), \qquad \mathrm{Bot}_{re}(K_n) = \Omega(1.1^n)$$

3.5 Tree Graph

Next, we investigate the Tree graph. In particular, we show that, for fixed p, q, the bottleneck complexity is logarithmic in n which matches the generic lower bound from Theorem 2. In communication complexity, the Tree graph is asymptotically optimal but the associated constant is worse than the one for Empty graph.

Leafs and Distances. A *leaf* of a Tree Graph is a vertex with outdegree 1 and indegree 0. A *distance* from vertex u to vertex v is the length of the shortest path connecting them, i.e. the minimum $k \in \mathbb{N}$ for which there exist vertices $u = v_0, v_1, \ldots, v_k = v$ such that $(v_{i-1}, v_i) \in E$ for all $i = 1, \ldots, k$.

Informal idea. A binary tree on n vertices contains roughly $n/2$ leaves, each having distance $\Theta(\log n)$ from the teacher. It can be shown that focusing on a single leaf, the expected number of rounds before this leaf is convinced is logarithmic in n. However, this is not enough to conclude that $\mathrm{Rnd}(T_n, \ell)$ is logarithmic. In order to draw the desired conclusion, we need a result stating that a single leaf is convinced in short time with high probability (not only that it is convinced in short time on average). A common tool to derive such concentration results for random variables is called a Chernoff Bound (see Proposition 4). We use Chernoff Bound in the proof of the following lemma.

Path graph. A *Path graph* is a directed graph P_n with $n+1$ vertices $v_0 = T$, $v_1, \ldots, v_n$ and n edges $v_i \to v_{i-1}$ for $i = 1, \ldots, n$.

Lemma 4. *Let P_n be a path. Then*

$$\left(\frac{1}{p} - \frac{1}{2}\right) \cdot n \leq \mathrm{Rnd}(P_n) \leq \frac{1}{p} \cdot n$$

and $\mathbb{P}[R(P_n) > 4/p \cdot n] < \frac{1}{3^n}$.

Proof. We first bound $\mathrm{Rnd}(P_n)$. Assume we already convinced first k learners (for some $k = 0, \ldots, n-1$). Denote by p_i the probability that in the next round we convince at least i more learners. Then for $i = 1, \ldots, n-k$ we have

$$p_i \geq \frac{1}{i!} \cdot p^i,$$

since for that to happen, it suffices if the i edges (vertices) appear within the round in one particular relative ordering out of $i!$ possible orderings and all i used edges result in convincing the next learner (which has probability p each).

The expected number of learners convinced in this round is then $S = \sum_{i=1}^{n-k} p_i$. Using the well known series $e^t = \sum_{i=0}^{\infty} t^i/i!$ we get

$$p = p_1 \leq S \leq \sum_{i=1}^{\infty} \frac{1}{i!} p^i = e^p - 1.$$

Since altogether we need to convince n learners, by linearity of expectation we have

$$\frac{1}{e^p - 1} \cdot n \leq \mathrm{Rnd}(P_n) \leq \frac{1}{p} \cdot n.$$

It remains to check that the left-hand side is greater than or equal to $(1/p - 1/2) \cdot n$. This is easily done by performing Taylor expansion of

$$(1/p - 1/2) \cdot (e^p - 1) = (1/p - 1/2) \left(\sum_{i=1}^{\infty} \frac{1}{i!} p^i \right) = 1 - \sum_{i=2}^{\infty} \frac{i-1}{2i!} p^i < 1.$$

Now we turn to the concentration result. Note that until we are done, the probability that we convince at least one more learner in a round is at least p. For $i = 1, \ldots, 4/p \cdot n$, let X_i be a random variable taking value 1 with probability p and value 0 otherwise, independently on the other variables, and let $X = \sum_{i=1}^{4/p \cdot n} X_i$. The variables X_i indicate if we succeeded in convincing a learner in i-th round. If at least n of them are equal to one, we are done, so we want to upper bound $\mathbb{P}[X < n]$. We have $\mathbb{E}[X] = p \cdot 4/p \cdot n = 4n$. By Chernoff Bound (see Proposition 4),

$$\mathbb{P}[R(P_n) > 4/p \cdot n] = \mathbb{P}[X < n] = \mathbb{P}[X < (1 - 3/4) \cdot \mathbb{E}[X]] < e^{-4n\frac{(\frac{3}{4})^2}{2}} = e^{-\frac{9}{8}n} < \frac{1}{3^n}.$$

□

Theorem 4. *Let T_n be a tree graph with $n+1$ nodes. Then*

$$\left(\frac{1}{p} - \frac{1}{2}\right) \log n \leq \mathrm{Rnd}(T_n) \leq \frac{8}{p} \cdot \log n.$$

Proof. Divide the process into phases of $\log n \cdot 4/p$ rounds each. Pick an arbitrary vertex v. By Lemma 4, the probability that it is not convinced in a single phase is at most $r = \frac{1}{3^{\log_2 n}}$. By Union Bound (see Proposition 2), the probability that we finish within a single phase is at least $1 - n \cdot r \geq 1 - 1/\sqrt{n}$, which is at least $1/2$ for $n \geq 4$. Hence for $n \geq 4$ the expected number of phases is at most two implying that $\mathrm{Rnd}(T_n) \leq 2 \cdot 4/p \cdot \log n = 8/p \cdot \log n$. □

As an immediate consequence, we obtain the fourth line of Table 1.

Corollary 3. *For (p, q)-learners we have*

$$\mathrm{Rnd}(T_n) = \Theta(T \cdot \log n), \qquad \mathrm{Comm}(T_n) = \Theta(T \cdot n \log n), \qquad \mathrm{Bot}(T_n) = \Theta(T \cdot \log n).$$

3.6 Layered Hierarchy Graphs

Finally, we discuss various Hierarchy graphs. As compared to the Empty graph, Layered Hierarchy graphs achieve exponentially better rounds complexity while (asymptotically) matching the communication complexity. The idea behind the 2-Hierarchy is to specify a small group of learners who, once they all acquire teacher's language, will never lose it again. Convincing this smaller group will be fast and all the learners from the group can then serve as additional teachers, speeding up the process of convincing the remaining learners. In Hierarchy, this idea is iterated (applied recursively) to obtain a sequence of groups, quickly increasing in size, each of which will be quickly convinced once the previous group is convinced.

Construction of Layered Hierarchies. Given a finite sequence of integers $s = (s_1, \ldots, s_k)$, a *Layered Hierarchy* determined by s is the following graph $\mathrm{H}(s) = \mathrm{H}(s_1, s_2, \ldots, s_k)$:

- Vertices: For each $i = 1, \ldots, k$ define S_i as a set of s_i vertices. Also, set $S_0 = \{T\}$ to be the teacher. Hence we get $1 + \sum_{i=1}^{k} s_i$ vertices.
- Edges: For every $i = 1, \ldots, k$, every $u \in S_{i+1}$ and every $v \in S_i$ include a directed edge $u \to v$.

Of particular interest are Layered Hierarchies with each group at least exponentially larger than the previous one. It's readily checked that in such cases, the resulting Layered Hierarchy graph has $\Theta(s_k)$ vertices, $\Theta(s_k \cdot s_{k-1})$ edges, and maximum indegree $\Theta(s_k)$.

3.6.1 2-Hierarchy

Given a positive integer n, a *2-Hierarchy with n sinks* is a Layered Hierarchy graph

$$\mathrm{H}_n = \mathrm{H}(\log n/\log\log n, n).$$

Theorem 5. *Let H_n be a 2-Hierarchy. Then*

$$\mathrm{Rnd}(\mathrm{H}_n) = \frac{1}{p}\cdot\Theta(\log\log n).$$

Proof. Denote by T_1 the expected number of rounds to convince S_1 and by $T_{2|1}$ the expected number of rounds to convince S_2 once S_1 is convinced. Since no edges lead from S_1 to S_2, by linearity of expectation we have

$$T_{2|1} \le \mathrm{Rnd}(\mathrm{H}_n) \le T_1 + T_{2|1}.$$

By Theorem 1, $T_1 = \Theta(\log(\log n/\log\log n)) = O(\log\log n)$. Now assume all S_1 is already convinced. Learners in S_2 don't interact and within a single round, each of them listens to $|S_1|$ speakers who already speak teacher's language. Hence the second phase is $|S_1|$-times faster than teaching a classroom with n learners and one teacher. Theorem 1 yields

$$T_{2|1} = \mathrm{Rnd}(E_n)\cdot\frac{\log\log n}{\log n} = \frac{1}{p}\cdot\Theta(\log\log n).$$

Altogether, $\mathrm{Rnd}(\mathrm{H}_n) = \frac{1}{p}\cdot\Theta\left(\log\log n\right)$. □

The results concerning 2-Hierarchy are summarized in the following corollary which proves the fifth line of Table 1.

Corollary 4. *For (p,q)-learners we have*

$$\mathrm{Rnd}(\mathrm{H}_n) = \Theta(T\cdot\log\log n), \qquad \mathrm{Comm}(\mathrm{H}_n) = \Theta(T\cdot n\log n), \qquad \mathrm{Bot}(\mathrm{H}_n) = \Theta(T\cdot n\log\log n).$$

3.6.2 Hierarchy

Here we recursively apply the same idea to obtain a family of graphs GH_n such that $\mathrm{Rnd}(\mathrm{GH}_n) = \Theta(\log^\star n)$ while $\mathrm{Comm}(\mathrm{GH}_n) = \Theta(n\log n)$.

Construction of Hierarchy. Given a positive integer n, set $k = \log^\star n$ and construct a sequence $s = (s_1,\ldots,s_k)$ as follows:

- $s_k = n$,
- for every $i = k-1,\ldots,1$, set $s_i = \log s_{i+1}$,
- finally, reset s_{k-1} from $\log n$ to $\log n/\log^\star n$.

A *Hierarchy with n sinks* is a Layered Hierarchy graph $\mathrm{GH}_n = \mathrm{H}(s_1,\ldots,s_k)$. Note that by construction, $\log s_1 < 1$, hence for every $i = 1,\ldots,k-2$ we have $s_i \ge \log s_{i+1}$. However, $s_{k-1} < \log s_k$.

Theorem 6. *Let GH_n be a Hierarchy. Then*

$$\mathrm{Rnd}(\mathrm{GH}_n) = \frac{1}{p}\cdot\Theta\left(\log^\star n\right).$$

Proof. Set $k = \log^\star n$.

As in the proof of the 2-Hierarchy, for each $i = 1, \ldots, k$, define $T_{i|i-1}$ as the expected number of rounds to convince S_i assuming that the whole S_{i-1} is already convinced (in particular, $T_{1|0}$ is the expected time to convince S_1). Note that for $i = 1, \ldots, k-1$ we have $s_{i-1} \geq \log(s_i)$, hence $T_{i|i-1} = \Theta(1/p)$. For $T_{k|k-1}$, we get

$$T_{k|k-1} = \frac{\mathrm{Rnd}(E_n)}{s_{k-1}} = \frac{1}{p} \cdot \Theta\left(\log n \cdot \frac{\log^\star n}{\log n}\right) = \frac{1}{p} \cdot \Theta\left(\log^\star n\right).$$

Overall,

$$\mathrm{Rnd}(\mathrm{GH}_n) \leq \sum_{i=1}^{k} T_{i|i-1} = (\log^\star n - 1) \cdot \Theta(1/p) + \Theta(\log^\star n),$$

and

$$\mathrm{Rnd}(\mathrm{GH}_n) \geq T_{k|k-1} = \frac{1}{p} \cdot \Theta\left(\log^\star n\right)$$

hence $\mathrm{Rnd}(\mathrm{GH}_n) = \frac{1}{p} \cdot \Theta\left(\log^\star n\right)$. □

The results concerning Hierarchy are summarized in the following corollary which yields the sixth line of Table 1.

Corollary 5. *For (p, q)-learners we have*

$$\mathrm{Rnd}(\mathrm{GH}_n) = \Theta(T \cdot \log^\star n), \qquad \mathrm{Comm}(\mathrm{GH}_n) = \Theta(T \cdot n \log n), \qquad \mathrm{Bot}(\mathrm{GH}_n) = \Theta(T \cdot n \log^\star n).$$

3.7 Non-symmetrically overlapping languages

Here we discuss how our results on (p, q)-learning yield a bound for the scenario with non-symmetric overlaps among languages.

Traditionally, memoryless learning has been studied in the setting of ℓ languages $L_1, \ldots, L_\ell$ that are allowed to overlap. For every pair of languages L_i, L_j, let q_{ij} be the probability that a listener with language L_i will successfully parse a sentence generated by a speaker with language L_j. Hence $q_{ii} = 1$ for all $i = 1, 2, \ldots, \ell$ and in general $q_{ij} \neq q_{ji}$.

A memoryless learner who has just received a sentence that she cannot parse considers switching to another language. In general, this new language is determined by a probability distribution on all ℓ languages. For $i, j, k = 1, 2, \ldots, \ell$, let p_{ijk} be the probability that a learner with hypothesis L_i, having heard a sample sentence from an individual with language L_j, decided to switch to L_k. Thus, in the standard scenario with learners who switch to a language selected uniformly at random among the remaining $\ell - 1$ languages, for all $i = 1, 2, \ldots, \ell$ we would have $p_{iii} = 1$, $p_{iij} = 0$ for $j \neq i$, and $p_{ijk} = (1 - q_{ij}) \cdot 1/(\ell - 1)$ for $j \neq i$ and all $k \neq i$.

Let G be any of the considered communication graphs (an empty graph, a tree graph, or any Layered hierarchy graph) with n memoryless learners. Let $L_1, \ldots, L_\ell$ be a set of overlapping languages and let $P = (p_{ijk})^{\ell \times \ell \times \ell}$ store all the values p_{ijk} for $i, j, k = 1, 2, \ldots, \ell$ defined as above. Denote by $\mathrm{Rnd}(G, P)$ the rounds complexity of the learning process on G with n memoryless learners behaving according to P.

Denote by $p_{\min} = \min\{p_{ij1} \mid i, j = 1, 2, \ldots, \ell,\ i \neq j\}$ the smallest probability of switching to the right language, after listening to one sample sentence from an individual with a different language. It's straightforward to check that all the upper bounds presented in Section 3 so far apply with p replaced by $p_{\min}$. These results are summarized in the following corollary.

Corollary 6. *Let E_n, T_n, H_n, GH_n be an empty graph, a tree graph, a 2-Hierarchy graph and a Hierarchy graph with n learners, respectively. For $i, j, k = 1, 2, \ldots, \ell$, let p_{ijk} be the probability that a learner with hypothesis L_i, having heard a sample sentence from an individual with language L_j, decided to switch to L_k. Let $p_{\min} = \min\{p_{ij1} \mid i, j = 1, 2, \ldots, \ell,\ i \neq j\}$. Then*

$$\mathrm{Rnd}(E_n, P) = O(1/p_{\min} \cdot \log n), \qquad \mathrm{Rnd}(\mathrm{H}_n, P) = O(1/p_{\min} \cdot \log \log n),$$
$$\mathrm{Rnd}(T_n, P) = O(1/p_{\min} \cdot \log n), \qquad \mathrm{Rnd}(\mathrm{GH}_n, P) = O(1/p_{\min} \cdot \log^\star n).$$

4 Batch learning

In this section we analyze the efficiency of batch learners with ℓ languages and symmetric overlap q on various graph families. In the setting with multiple learners we investigate the dependency on n and ℓ for fixed q.

4.1 Single teacher, single learner

First, we consider the baseline setting of a single teacher and single learner. We compute the expected number T of sample sentences required to convergence for a single batch learner. Later on, we will use this result to express the efficiency of the learning process in the scenario with n-learners in terms of n and T.

Lemma 5. *Let q, ℓ be fixed. Then*

$$\left(1 - \frac{1}{e}\right) \cdot \log_{1/q}(\ell - 1) \leq T \leq \log_{1/q}(\ell - 1) + \frac{1}{1-q},$$

hence for fixed q and $\ell \to \infty$ we get $T = \Theta(\log \ell)$.

Proof. We proceed as in the proof of Theorem 1. Informally, for each of the $\ell - 1$ "wrong" languages, the learner needs to receive at least one sample sentence that is not consistent with it. Since, on average, each sample sentence is consistent with only a constant fraction (about q-portion) of the languages, we expect to be done in approximately $\log \ell$ steps.

Formally, for $k = 0, 1, \ldots$ let w_k be a random variable equal to the number of non-teacher languages that are consistent with the first k sample sentences. Then $T = \sum_{k=0}^{\infty} \mathbb{P}[w_k > 0]$. Clearly we have $\mathbb{E}[w_0] = w_0 = \ell - 1$ and by linearity of expectation we obtain $\mathbb{E}[w_{k+1}] = \mathbb{E}[w_k] \cdot q$. Proceeding as in the proof of Theorem 1 we get the same result with n replaced by $\ell - 1$ and $1 - p$ replaced by q. □

4.2 Empty graph

Theorem 7. *Let E_n be the Empty graph on $n + 1$ vertices. Then*

$$\left(1 - \frac{1}{e}\right) \cdot \log_{1/q}(n(\ell - 1)) \leq \mathrm{Rnd}(E_n) \leq \log_{1/q}(n(\ell - 1)) + \frac{1}{1-q},$$

hence for fixed q we get $\mathrm{Rnd}(E_n) = \Theta(T + \log n)$.

Proof. This is an easy consequence of Lemma 5. Informally, each of the $\ell - 1$ "wrong" languages has to be excluded for each of the n learners. Since all these exclusion events are independent and have the same constant probability q, we expect to be done in approximately $\log(\ell \cdot n)$ steps.

Formally, we define random variables $w_0, w_1, \ldots$ as in the proof of Lemma 5 and obtain the same bounds with $\ell - 1$ replaced by $n(\ell - 1)$. □

Corollary 7. *For batch learners with symmetric overlap q we have*

$$\mathrm{Rnd}(E_n) = \Theta(T + \log n), \qquad \mathrm{Comm}(E_n) = \Theta(T \cdot n + n \log n), \qquad \mathrm{Bot}(E_n) = \Theta(T \cdot n + n \log n).$$

4.3 Complete graph

Here we briefly study the complete graph. Recall that the rounds complexity for (p, q)-learners on a complete graph is exponential in n. For batch learners, the process might not even converge with probability one. Intuitively, batch learners can form a group that shares a wrong language hypothesis and, by listening to each other, reinforce this wrong hypothesis strongly enough that it will never be convinced by the teacher. The following lemma shows one such toy example.

Lemma 6. *Consider a complete graph with one teacher and three batch learners. Assume that $q < 0.5$ and $\ell = 2$. Then, with positive probability, the learning process will not converge to all learners speaking the teacher's language.*

Proof. Assume that in the first round, all the sentences spoken by the teacher were consistent with L_2 while no sentences spoken by the learners were consistent with L_1. This event has a fixed positive probability $q^3(1-q)^6$ and the first round then results in each learner having heard 3 sentences consistent with L_2 and only one sentence consistent with L_1.

For each learner i ($i = 1, 2, 3$), let x_i^k be a random variable denoting the difference between the number of received sentences consistent with L_2 and those consistent with L_1, after k rounds. As noted above, $x_1^1 = x_2^1 = x_3^1 = 3 - 1 = 2$ with probability $q^3(1-q)^6$.

If $x_i^k \geq 2$ for all $i = 1, 2, 3$ and $k = 1, 2, \ldots$ then no learner ever switches to L_1. Assume that this is not the case and let t be the smallest time such that $\min\{x_1^t, x_2^t, x_3^t\} = 1$. Without loss of generality, $x_1^t = 1$.

Up till the time-point t, the first learner listened to one sentence from the teacher (with hypothesis L_1) and two sentences from other learners (with hypotheses L_2). A net outcome of the round is the difference $x_1^{k+1} - x_1^k$. This difference is negative only if the teacher's sentence was not consistent with L_2 but both learners' sentences were consistent with L_1, that is with probability $p^- = (1-q)q^2$. On the other hand, the difference is positive in several cases, including the one when each sentence is only consistent with its language (probability $p^+ > (1-q)^3$). Since $p^-/p^+ < (q/(1-q))^2 < 1$, the sequence of random variables $\{x_1^k\}_{k=1}^t$ is a biased random walk starting at value 2. The probability p that such a biased random walk always stays greater than 1 is positive. Therefore with probability at least $q^3(1-q)^6 \cdot p > 0$ we have $x_i^k \geq 2$ for all $i = 1, 2, 3$ and $k = 1, 2, \ldots$, that is none of the learners ever switches to L_1 and the process doesn't converge to all the learners speaking the teacher's language. □

4.4 Tree graph

Here we analyze the tree graph. It turns out that batch learners with symmetric overlap are less efficient than (p, q)-learners. Intuitively, this is due to the fact that the tree is deep. By the time the teacher's language propagates to the bottom level of the tree, the nodes there have already

listened to many sentences from some wrong language. Since they remember all of them, making them switch to the teacher's language will take long.

As in the case of the (p, q)-learners, we first analyze the path graph.

Lemma 7. *Let P_n be a path and let q, ℓ be fixed. Then*

$$\mathrm{Rnd}(P_n) \geq 2^{n-1} \cdot T.$$

Proof. We proceed by mathematical induction. Clearly $\mathrm{Rnd}(P_1) = T = 2^{1-1} \cdot T$. Now assume the claim is true for $n-1$. Denote by t_{n-1} the time point when we have just convinced the first $n-1$ learners in the path. By induction, $\mathbb{E}[t_{n-1}] = 2^{n-2} \cdot T$. By that time, the expected number of sample sentences heard by the last learner is also $\mathbb{E}[t_{n-1}] = 2^{n-2} \cdot T$ and due to the initial condition, all these sample sentences were spoken by an individual with hypothesis L_ℓ. Hence the last learner heard, on average, $t_{n-1} \cdot (1-q)$ more sentences consistent with L_ℓ than with L_0. By symmetry between L_0 and L_ℓ, in order to catch up with this head start, we now need to provide the last learner with at least t_{n-1} sentences sampled from an individual with hypothesis L_0. By linearity of expectation, $t_n \geq t_{n-1} + t_{n-1} = 2^{n-1}$ and we are done. □

Theorem 8. *Let T_n be a tree graph with $n+1$ nodes. Then*

$$\mathrm{Rnd}(T_n) = \Omega(T \cdot n).$$

Proof. A tree graph with $n+1$ nodes has $\Omega(\log n)$ layers, hence it contains at least one path of length $\Omega(\log n)$. By Lemma 7, converging on this path only takes time $\Omega(2^{\log n - 1} \cdot T) = \Omega(n \cdot T)$, hence converging on the whole graph takes $\Omega(n \cdot T)$. □

Corollary 8. *For batch learners with symmetric overlap q we have*

$$\mathrm{Rnd}(T_n) = \Omega(T \cdot n), \qquad \mathrm{Comm}(T_n) = \Omega(T \cdot n^2), \qquad \mathrm{Bot}(T_n) = \Omega(T \cdot n).$$

4.5 2-Hierarchy

Here we analyze 2-Hierarchy graph. Intuitively, we argue that first layer is usually convinced quickly and once that happens, convincing the second layer happens quickly too. We use Chernoff bound. We note that we don't optimize constants and focus on the main idea of the argument.

Theorem 9. *Let H_n be a 2-Hierarchy graph and let $q \leq 0.25$ be fixed. Then*

$$\mathrm{Rnd}(H_n) = \Theta(T + \log\log n).$$

Proof. For the lower bound, by Theorem 7 convincing just the first layer S_1 takes

$$\Theta(T + \log|S_1|) = \Theta\left(T + \log\left(\frac{\log n}{\log\log n}\right)\right) = \Theta(T + \log\log n - \log\log\log n) = \Theta(T + \log\log n).$$

For the upper bound, we proceed in two stages. (1) First, we show that, for some constant c_0 independent of n and T, the probability that the first layer is not convinced after $i \cdot c_0 + \Theta(T + \log\log n)$ rounds is at most 2^{-i}; (2) Second, given that the first layer was convinced in r rounds, we show that the probability that the second layer will not be convinced after $2 \cdot (j+1) \cdot r$ more rounds is at most 2^{-j}.

Point (2) implies that, in expectation, convincing the second layer takes at most $r \cdot \sum_{j=1}^{\infty} 2(j+1)/2^{j-1} = 12 \cdot r$ more rounds. Point (1) implies that, in expectation, r is at most $2 \cdot c_0 + \Theta(T + \log\log n) = \Theta(T + \log\log n)$. Altogether, this implies that, in expectation, the whole process takes $\Theta(T + \log\log n)$ rounds.

(1) For $k \geq 0$, let w_k be the random variable that denotes the number of wrong languages that are not yet excluded from consideration among individuals in the first layer, after k sample sentences (including repetitions). We have $w_0 = (\ell-1)\cdot|S_1|$ and as in the proof of Theorem 1, $\mathbb{E}[w_k] = (\ell-1)\cdot|S_1|\cdot q^k$.

Fix an integer i and consider $r = \log_{1/q}(2^i \cdot (\ell-1)\cdot\log n) = i\cdot\log_{1/q}(2) + \Theta(T + \log\log n)$ rounds (that is, we set $c_0 = \log_{1/q} 2$). Then $\mathbb{E}[w_r] \leq 1/2^i$, hence by Markov inequality $\mathbb{P}[w_r > 0] = \mathbb{P}[w_r \geq 1] \leq 2^{-i}$.

(2) Recall that $T = \Theta(\ell)$. Assume the first layer was convinced in $r_1 \geq 3(\log\ell + \log\log n)$ rounds (if not, wait until then), fix an integer $j > 0$ and consider $r_2 = 2\cdot(j+1)\cdot r_1$ more rounds. Fix a learner A from the second layer and one wrong language, say L_1. We claim that the probability that after $r_1 + r_2$ rounds individual A still favours L_1 over the correct L_0 is small.

In the first r_1 rounds, A might have heard up to $r_1\cdot|S_1|$ sentences consistent with L_1 and perhaps no sentence consistent with the correct L_0. In the next r_2 rounds, A heard $r_2\cdot|S_1|$ sentences consistent with the correct L_0. Let $X_1,\ldots,X_{r_2\cdot|S_1|}$ be independent random variables indicating if those sentences were also consistent with L_1. That is, for each $i = 1,\ldots,r_2\cdot|S_1|$ we have

$$X_i = \begin{cases} 1 & \text{with probability } q, \\ 0 & \text{with probability } 1-q. \end{cases}$$

Let $X = \sum_{i=1}^{r_2\cdot|S_1|} X_i$ be the sum of all these random variables and let

$$\mu = \mathbb{E}[X] = r_2\cdot|S_1|\cdot q$$

be its expected value. Note that rewriting μ in terms of ℓ and n we get

$$\mu \geq 2\cdot(j+1)\cdot 3(\log\ell + \log\log n)\cdot\frac{\log n}{\log\log n}\cdot q \geq 6q(j+1)(\log\ell + \log n).$$

The probability p that A favours L_1 over L_0 is at most

$$p \leq \mathbb{P}[X \geq (r_2 - r_1)\cdot|S_1|].$$

We rewrite

$$\frac{(r_2-r_1)\cdot|S_1|}{\mu} = \frac{2j+1}{2(j+1)\cdot q} \geq 1 + \frac{1}{2q},$$

where the last inequality can be reduced to $q \leq 1/4 \leq j/(2j+2)$. Hence we can set $\delta = 1/(2q) > 1$ in Chernoff bound to get

$$p \leq \mathbb{P}[X \geq (r_2-r_1)\cdot|S_1|] \leq \mathbb{P}[X \geq (1+\delta)\cdot\mu] \leq e^{-\frac{1}{3}\delta\mu}.$$

We bound the right-hand side as

$$p \leq \exp\left(-\frac{1}{3}\delta\mu\right) \leq \exp\left(-\frac{1}{6q}\cdot 6q(j+1)(\log\ell+\log n)\right) = n^{-j-1}\cdot\ell^{-j-1}.$$

By Union Bound (see Proposition 2), the probability that some of the n learners in the second layer prefers some of the $\ell-1$ wrong languages is at most $p\cdot n\cdot(\ell-1) \leq (n\ell)^{-j} \leq 2^{-j}$ as desired.

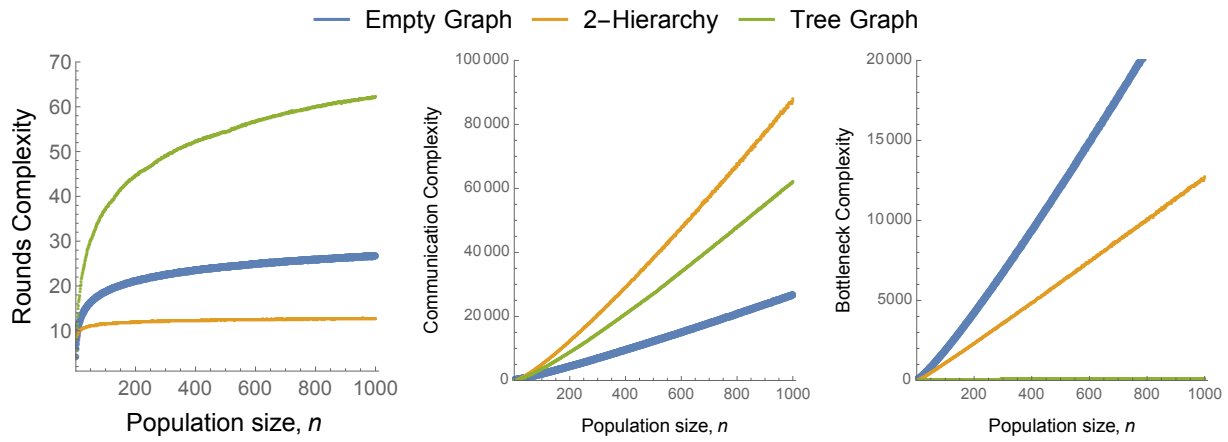

Figure SI.2: **Main graphs, all measures.** Rounds complexity, Communication complexity, and Bottleneck complexity against population size n for (p, q)-learners on Empty graph (blue), 2-Hierarchy (orange), and Tree graph (green). We set $\ell = 4$ and we consider a neutral teacher $p = q = 1/\ell$. All results are averages over 10 000 trials.

We conclude the proof. □

Corollary 9. *For batch learners with symmetric overlap q we have*

$$\mathrm{Rnd}(H_n) = \Theta(T + \log\log n), \quad \mathrm{Comm}(H_n) = \Theta\left(T \cdot \frac{n \cdot \log n}{\log\log n} + n \log n\right), \quad \mathrm{Bot}(H_n) = \Theta(T \cdot n + n \log\log n).$$

5 Additional simulations results

In this section we present extensive computer simulation results.

5.1 All complexity measures

Here we present simulation results for Rounds complexity, Communication complexity, and Bottleneck complexity on Empty graph, Tree graph, and 2-Hierarchy graph. We consider (p, q)-learners. Since the dependence on ℓ is linear, we fix $\ell = 4$ and consider a neutral teacher with $p = q = 1/\ell$.

We observe that in Rounds complexity, 2-Hierarchy is asymptotically better than both Empty graph and Tree graph. In Communication complexity, all three graphs are asymptotically equivalent and the Empty graph has the best associated constant. In Bottleneck complexity, the Tree graph achieves theoretical lower bound while 2-Hierarchy is worse and the Empty graph even worse.

5.2 Layered Hierarchies

Here we present simulations comparing various Layered Hierarchy graphs. For simplicity we consider (p, q)-learners with neutral teacher $(p = q = 1/\ell)$ and we focus on 3-Layered Hierarchies and compare the following three graphs with the baseline Empty graph:

- *Const-Hierarchy* is a 3-Layered Hierarchy with $n/3$ learners in each layer, i.e. the sizes of the layers are constant.

	Rounds Compl.	Communication Compl.	Bottleneck Compl.
Empty graph	$\Theta(\log n)$	$\Theta(n \log n)$	$\Theta(n \log n)$
Const-Hierarchy	$\Theta(\log n)$	$\Theta(n^2 \log n)$	$\Theta(n \log n)$
Lin-Hierarchy	$\Theta(\log n)$	$\Theta(n^2 \log n)$	$\Theta(n \log n)$
Exp-Hierarchy	$\Theta(\log \log \log n)$	$\Theta(n \log n)$	$\Theta(n \log \log \log n)$

Table 3: **Asymptotic complexity measures for 3-Layered Hierarchies.** The complexity measures for (p, q)-learners on different 3-Layered Hierarchies, as a function of population size n (recall that the dependence on the number of languages ℓ is always linear). The first line shows the Empty graph for reference. The second line shows the Const-Hierarchy, i.e. a graph consisting of 3 layers of size $n/3$ each. The third line shows the Lin-Hierarchy, i.e. a graph with layers of size $n/6$, $n/3$, $n/2$, respectively. The third line shows the Exp-Hierarchy, i.e. a graph with n learners and first two layers of size $\log \log n$ and $\log n / \log \log n$. Note that as compared to the Empty graph, the Exp-Hierarchy asymptotically improves both the Rounds complexity and the Bottleneck complexity while the other two Layered Hierarchies do not.

- *Lin-Hierarchy* is a 3-Layered Hierarchy with $n/6$, $n/3$, and $n/2$ learners in the respective layers, i.e. the sizes of the layers increase linearly.
- *Exp-Hierarchy* is a 3-Layered Hierarchy with first two layers containing $\log \log n$ and $\log n / \log \log n$ learners, i.e. the Exp-Hierarchy is a 3-layered analogue of Hierarchy GH_n.

The asymptotic Rounds complexity, Communication complexity, and Bottleneck complexity for Empty graph, and for Const-, Lin-, and Exp-Hierarchy are summarized in Table 3.

We present results of two different simulation experiments. The first one illustrates that dependency on ℓ, the second one illustrates dependency on n.

Illustration of dependency on ℓ. In this experiment we fix the number of learners n ($n = 120$, 240, 480) and consider the dependency on ℓ. Figure SI.3 shows that this dependence is linear in all cases which matches the theoretical results of Corollaries 1 and 5. We summarize the results in more detail.

- *Rounds complexity.* As stated in the main article, Const- and Lin-Hierarchy do not improve the asymptotic rounds complexity on n as compared to the Empty graph. However, for fixed n, they still improve the rounds complexity by a constant factor. The Exp-Hierarchy gives strict asymptotic improvement that shows even for small n. See the first row of Figure SI.3.
- *Communication complexity.* Since both Const- and Lin-Hierarchy contain $\Theta(n^2)$ edges, their communication complexity is at least quadratic (asymptotically worse than that of the Empty graph) and this shows even for small n. The Exp-Hierarchy matches the asymptotic complexity of the Empty graph, although the associated constant is worse. See the second row of Figure SI.3.
- *Bottleneck complexity.* Recall that the Bottleneck complexity is the Rounds complexity times the maximum degree. The Const- and Lin-Hierarchy have the same asymptotic complexity as the Empty graph, however, the associated constants are better. The Exp-hierarchy improves the asymptotic complexity of the Empty graph, but for small population sizes the Const- and Lin-hierarchies are better than the Exp-Hierarchy. See the third row of Figure SI.3.

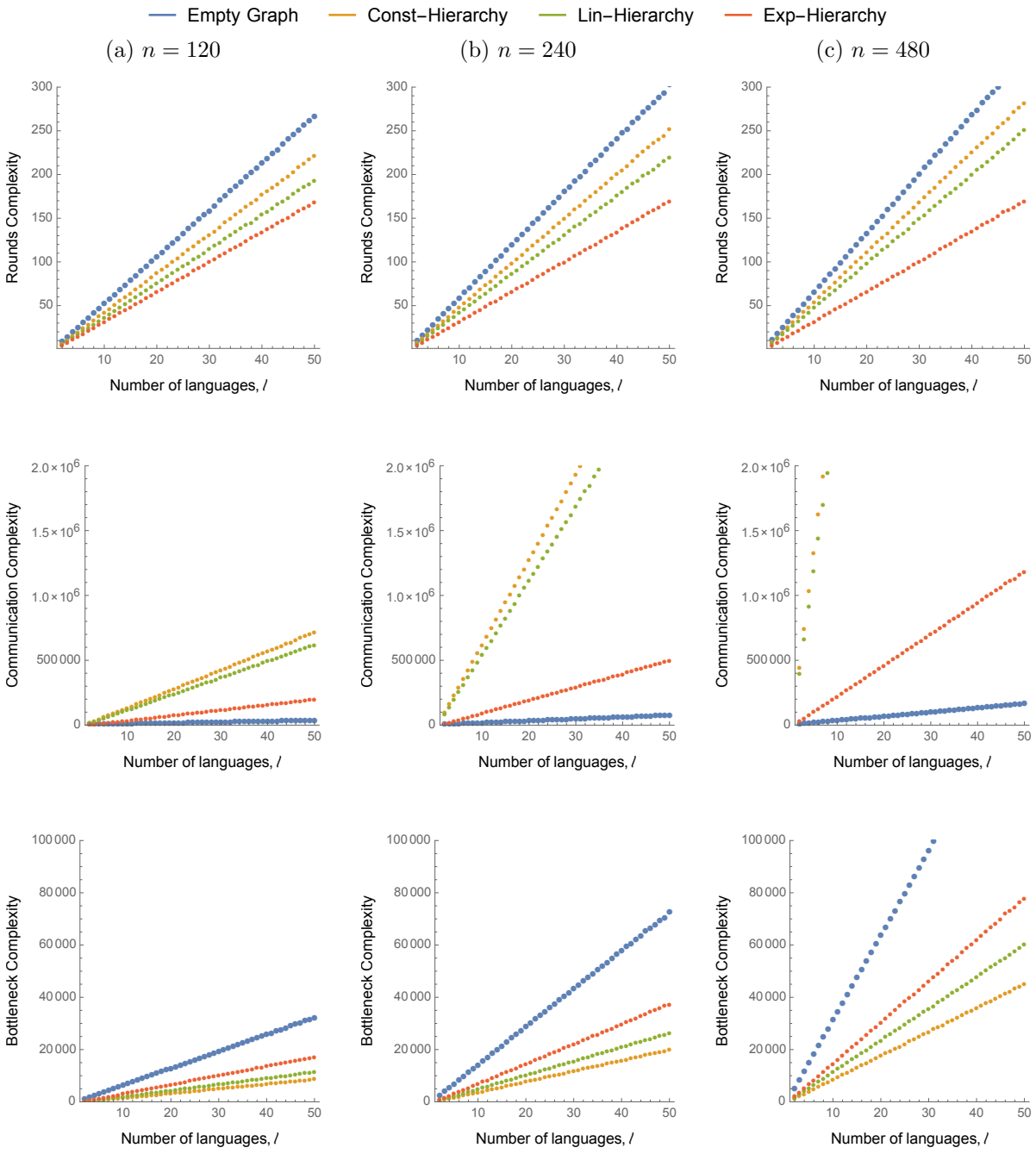

Figure SI.3: **Layered Hierarchies.** Simulation results for (p, q)-learners ($p = q = 1/\ell$) on various Layered Hierarchy graphs with x-axis representing the number of languages ($2 \leq \ell \leq 50$) and y-axis representing the Rounds complexity, Communication complexity, and Bottleneck complexity in the respective rows. The number of individuals is fixed to 120, 240, and 480, in the respective columns. All results are averages over 10 000 trials. The color representation is Blue: Empty graph; Orange: Const-Hierarchy $\mathrm{H}(n/3, n/3, n/3)$; Green: Lin-Hierarchy $\mathrm{H}(n/6, n/3, n/2)$; Red: Exp-Hierarchy $\mathrm{H}(\log\log n, \log n/\log\log n, n - \log n/\log\log n - \log\log n)$.

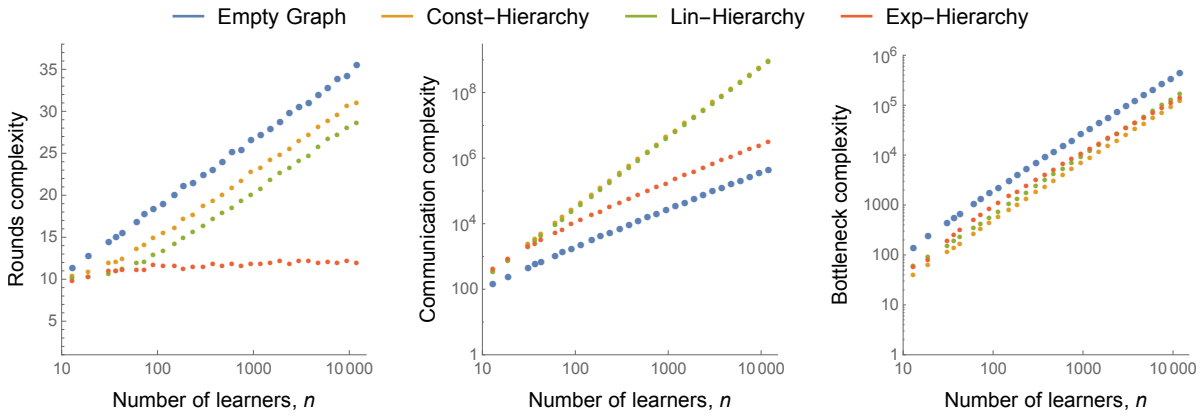

Figure SI.4: **Layered Hierarchies, dependence on n.** Simulation results for (p, q)-learners on various Layered Hierarchy graphs with x-axis representing the number of learners ($10 \leq n \leq 10\,000$) and y-axis representing the Rounds complexity, Communication complexity, and Bottleneck complexity, respectively. We set $\ell = 4$ and $p = q = 1/\ell$. All results are averages over 1 000 trials. All axes apart from y-axis in Rounds complexity are log scale. The color representation is Blue: Empty graph; Orange: Const-Hierarchy $\mathrm{H}(n/3, n/3, n/3)$; Green: Lin-Hierarchy $\mathrm{H}(n/6, n/3, n/2)$; Red: Exp-Hierarchy $\mathrm{H}(\log\log n, \log n, n - \log n - \log\log n)$.

Illustration of dependency on n. As shown in Figure SI.3, the dependency on ℓ is linear. Hence to illustrate the dependency on n, we can fix ℓ (we set $\ell = 4$) and present simulation results for large population sizes (see Figure SI.4).

We observe that for small population sizes, the Bottleneck complexity of Lin-Hierarchy is better than that of Exp-Hierarchy. However, Exp-Hierarchy is asymptotically better and this shows from $n \sim 3\,500$ on. Similarly, Exp-Hierarchy eventually outperforms Const-Hierarchy; namely when the Rounds complexity of Const-Hierarchy becomes three times larger than that of Exp-Hierarchy (see Figure SI.4).

5.3 Random sparse graphs

Here we present simulations for (p, q)-learners on random sparse graphs.

Given n learners and a positive real number $c \in [0, 1]$, we create a random graph $G(n, c)$ by including each of the n^2 edges of a Complete graph with the same probability c, independently of the other edges.

For large c, the resulting graphs are dense and behave very much like the complete graph. To obtain more interesting sparse graphs, we set $c = \log n/n$. By Theorem 5 in [3], this guarantees that for large n, the resulting graph satisfies the conditions of Lemma 1 with constant probability. Hence for fixed n we generate random graphs $G(n, \log n/n)$, check for the conditions of Lemma 1 and simulate the process.

The results show that random graphs are worse than the baseline comparison of Empty graph. Thus randomly chosen graphs do not improve the complexity measures, and finding graphs that improve the complexity measures is valuable. In this work we present the graph structures (e.g., Hierarchy graphs) which significantly improve the complexity bounds.

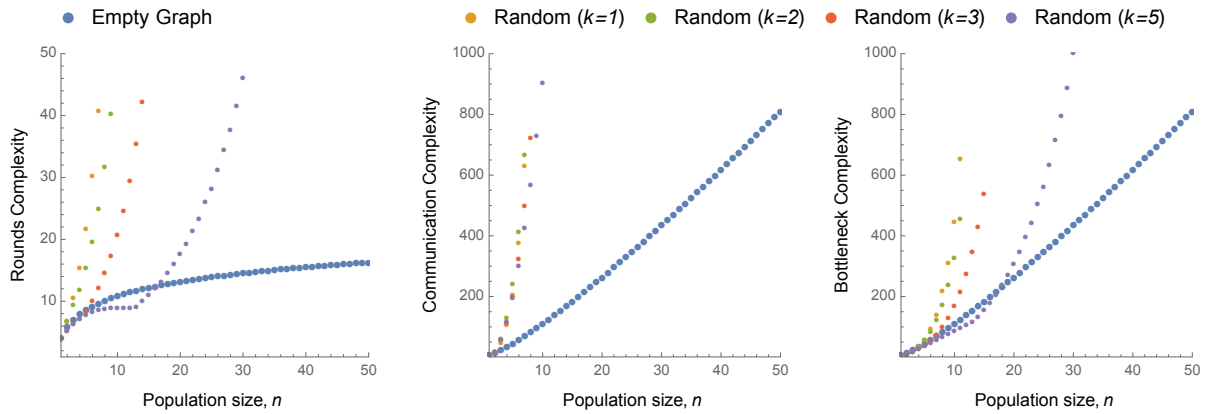

Figure SI.5: **Simulation results of random sparse graphs.** The x-axis represents the number n of learners, the y-axis represents Rounds complexity, Communication complexity, and Bottleneck complexity, respectively. Each dot is an average over 150 000 trials. For random graphs, we generate 1 000 random graphs and run 150 replicates on each. For the Empty graph, we run 150 000 replicates. As above we set $\ell = 4$ and $p = q = 1/\ell$. The color coding is as follows: Blue: Empty graph; Orange: $p = \log n/n$; Green: $p = 2 \cdot \log n/n$; Red: $p = 3 \cdot \log n/n$; Purple: $p = 5 \cdot \log n/n$. Since Empty graph is the baseline comparison, the blue color appears in bold font.

5.4 Distribution of the number of rounds

Here we present simulation results showing the full distribution of the number of rounds until the process converges (as opposed to showing the rounds complexity which is the *expected* number of rounds until the process converges).

We consider Empty graph, 2-Hierarchy and Tree graph with $n = 100$ learners. We fix $\ell = 10$ and consider both memoryless learners ($p = q = 1/\ell$) and batch learners (with symmetric overlap $q = 0.1$), see Figure SI.6(a),(b). Furthermore, we consider the case of non-symmetric overlap where language L_2 is quite similar to the teacher's language L_1 while all other languages are quite different. Again we consider both memoryless learners and batch learners (see Figure SI.6(c),(d))

In all four scenarios, 2-Hierarchy is faster than the Empty graph. For batch learners, Tree graph is much slower than both Empty graph and 2-Hierarchy so we don't show it.

6 Further Directions

In this work we present novel and natural extension of the traditional learning framework, and present answers to the main scientific question. Our new framework also leads to many interesting research questions to pursue. We mention some of them below:

1. We considered populations of either weak memoryless learners or powerful batch learners. Populations of other types of learners can be studied.

2. While we present several interesting graphs (such as Hierarchies), a natural question is to consider other classes of graphs and study the trade-off they provide with respect to the complexity measures.

3. Another interesting direction is to consider other relevant complexity measures and study the effect of graphs on the complexity measures.

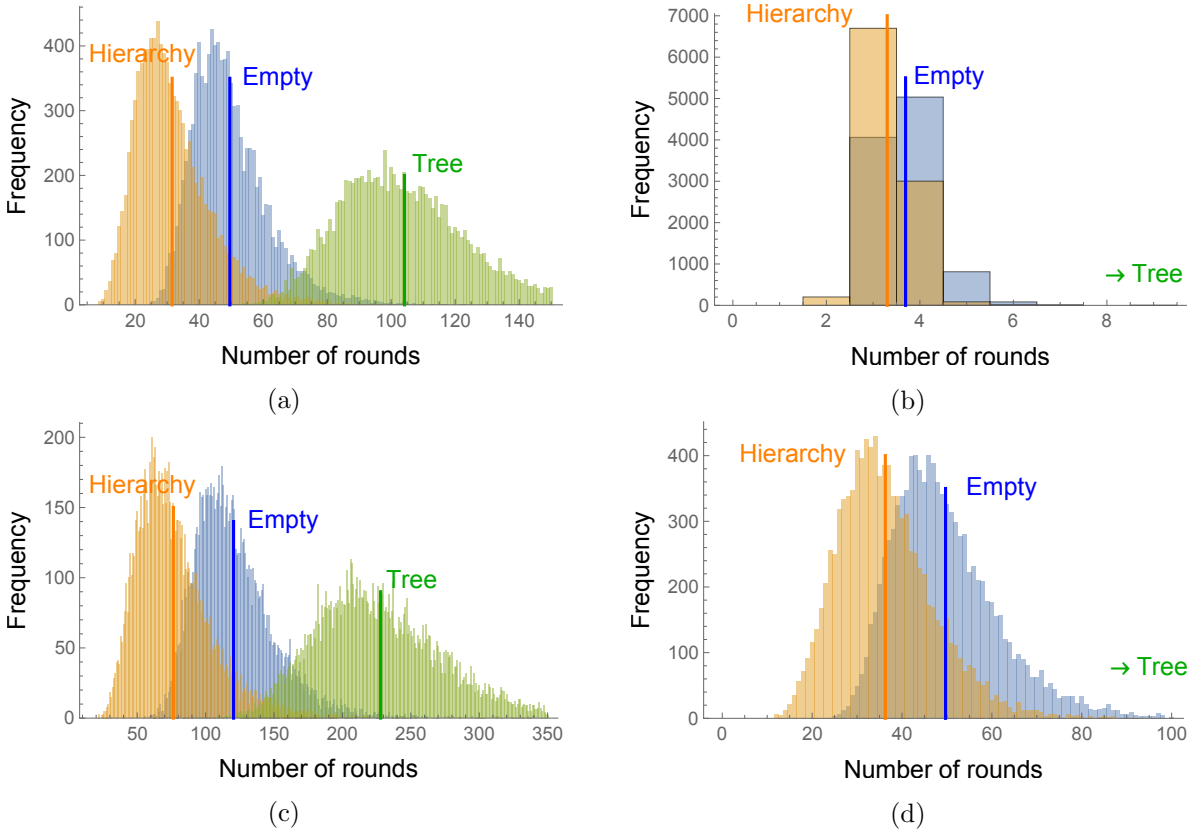

Figure SI.6: **Histograms for number of rounds.** We compare Empty graph (blue), 2-Hierarchy (orange), and Tree graph (green) for fixed population size $n = 100$ and fixed number of languages $\ell = 10$. For each graph we run the process 10 000 times. The x-axis represents the number of rounds until the process converges. The y-axis is the corresponding frequency. The colored line denotes the average (expected) number of rounds (i.e. the rounds complexity).

(a), (b). Symmetric overlap. In (a) we present results for memoryless learners (namely (p, q)-learners with $p = q = 1/\ell$). In (b) we present batch learners (symmetric overlap $q = 1/\ell = 0.1$).

(c), (d). Non-symmetric overlap. We set the relative overlap of L_1 and L_2 to 0.9 and all the other relative overlaps to 0.1. In (c), (d) we present results for memoryless learners and batch learners, respectively. For batch learners on Tree graph, the average number of rounds is approximately 80 and 1200 for symmetric and non-symmetric overlap, respectively. The intervals containing 90 % of the values are $[50, 120]$ and $[450, 2400]$, respectively.

4. Given a population size n what is the optimal structure with respect to some desired trade-offs for the complexity measures, and is there an algorithmic approach to solve the problem, are other interesting open questions.

References

[1] Motwani, R. & Raghavan, P. *Randomized algorithms* (Chapman & Hall/CRC, 2010).

[2] Karlin, S. *A first course in stochastic processes* (Academic press, 2014).

[3] Graham, A. J. & Pike, D. A. A note on thresholds and connectivity in random directed graphs. *Atl. Electron. J. Math* **3**, 1–5 (2008).